Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases

Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases

Yun R Li[1,2], Jin Li[1], Sihai D Zhao[3,43], Jonathan P Bradfield[1], Frank D Mentch[1], S Melkorka Maggadottir[1,4], Cuiping Hou[1], Debra J Abrams[1], Diana Chang[5,6], Feng Gao[5], Yiran Guo[1], Zhi Wei[7], John J Connolly[1], Christopher J Cardinale[1], Marina Bakay[1], Joseph T Glessner[1], Dong Li[1], Charlly Kao[1], Kelly A Thomas[1], Haijun Qiu[1], Rosetta M Chiavacci[1], Cecilia E Kim[1], Fengxiang Wang[1], James Snyder[1], Marylyn D Richie[8], Berit Flatø[9], Øystein Førre[9], Lee A Denson[10], Susan D Thompson[11], Mara L Becker[12], Stephen L Guthery[13], Anna Latiano[14], Elena Perez[15], Elena Resnick[16], Richard K Russell[17], David C Wilson[18], Mark S Silverberg[19], Vito Annese[20], Benedicte A Lie[21], Marilynn Punaro[22], Marla C Dubinsky[23], Dimitri S Monos[24,25], Caterina Strisciuglio[26], Annamaria Staiano[26], Erasmo Miele[26], Subra Kugathasan[27], Justine A Ellis[28,29], Jane E Munro[30,31], Kathleen E Sullivan[4,25], Carol Wise[32], Helen Chapel[33], Charlotte Cunningham-Rundles[16], Struan F A Grant[1,25], Jordan S Orange[34], Patrick M A Sleiman[1,25], Edward M Behrens[25,35], Anne M Griffiths[36], Jack Satsangi[37], Terri H Finkel[38], Alon Keinan[5,6], Eline T Luning Prak[39], Constantin Polychronakos[40], Robert N Baldassano[25,41], Hongzhe Li[39], Brendan J Keating[1,25] & Hakon Hakonarson[1,25,42]

Genome-wide association studies (GWASs) have identified hundreds of susceptibility genes, including shared associations across clinically distinct ~~disease groups and~~ autoimmune diseases. We performed an inverse χ^2 meta-analysis across ten pediatric-age-of-onset autoimmune diseases (pAIDs) in a case-control study including more than 6,035 cases and 10,718 shared population-based controls. We identified 27 genome-wide significant loci associated with one or more pAIDs, mapping to *in silico*–replicated autoimmune-associated genes (including *IL2RA*) and new candidate loci with established immunoregulatory functions such as *ADGRL2*, *TENM3*, *ANKRD30A*, *ADCY7* and *CD40LG*. The pAID-associated single-nucleotide polymorphisms (SNPs) were functionally enriched for deoxyribonuclease (DNase)-hypersensitivity sites, expression quantitative trait loci (eQTL), microRNA (miRNA)-binding sites and coding variants. We also identified biologically correlated, pAID-associated candidate gene sets on the basis of immune cell expression profiling and found evidence of genetic sharing. Network and protein-interaction analyses demonstrated converging roles for type 1, 2 and 17 helper T cell (T_H1, T_H2 and T_H17), JAK-STAT, interferon and interleukin signaling pathways in multiple autoimmune diseases.

Autoimmune diseases affect 7–10% of individuals living in the Western Hemisphere[1] and represent a significant cause of chronic morbidity and disability. High rates of familial clustering and comorbidity across autoimmune diseases suggest that genetic predisposition underlies disease susceptibility. GWASs and immune-focused fine-mapping studies of autoimmune thyroiditis (THY)[2], psoriasis (PSOR)[3], juvenile idiopathic arthritis (JIA)[4], primary biliary cirrhosis (PBC)[5], primary sclerosing cholangitis (PSC)[6], rheumatoid arthritis (RA)[7], celiac disease (CEL)[8], inflammatory bowel disease (IBD, which includes Crohn's disease (CD) and ulcerative colitis (UC)[9]), and multiple sclerosis (MS)[10,11] have identified hundreds of autoimmune disease–associated SNPs across the genome[12–14]. SNP associations in certain pan-autoimmune loci, such as *PTPN22* c.1858C>T (rs2476601), are evident in independent GWASs across multiple autoimmune diseases[15–18], whereas others have been uncovered through large-scale meta-analyses (for example, CEL-RA and type 1 diabetes (T1D)-CD) or by searches for known loci from one disease in another (for example, systemic lupus erythematosus (SLE))[19]. These studies demonstrate that more than half of genome-wide significant (GWS) autoimmune disease associations are shared by at least two distinct autoimmune diseases[20,21]. However, the degree to which common, shared genetic variations may similarly affect the risk of different pAIDs and whether these effects are heterogeneous have not been systematically examined at the genotype level across multiple diseases simultaneously.

RESULTS

Shared genetic risk associations across ten pediatric autoimmune diseases

We performed whole-genome imputation on a combined cohort of more than 6,035 pediatric subjects across 10 clinically distinct pAIDs (**Supplementary Table 1**) and 10,718 population-based control

A full list of affiliations appears at the end of the paper.

Received 22 February; accepted 23 July; published online XX XX 2015; doi:10.1038/nm.3933

subjects without prior history of autoimmune or immune-mediated disorders. We performed whole-chromosome phasing and used the 1,000 Genomes Project Phase I Integrated cosmopolitan reference panel (1KGP-RP) for imputation as previously described (SHAPEIT and IMPUTE2)[22,23]. Only individuals of self-reported European ancestry and confirmed by principal-component analysis (**Supplementary Fig. 1**) were included (Online Methods). Rare (minor allele frequency (MAF) < 1%) and poorly imputed (INFO score < 0.8) SNPs were removed, leaving a total of 7,347,414 variants.

Whole-genome case-control association testing was done using cases samples from each of the ten pAIDs and the shared controls, and additive logistic regression was applied with SNPTESTv2.5 (ref. 24). There was no evidence of genomic inflation. To identify shared pAID-association loci, we performed an inverse χ^2 meta-analysis, accounting for sample-size variation and the use of a shared control across the ten pAIDs[25]. We identified 27 linkage disequilibrium (LD)-independent loci, consisting of associated SNPs with $r^2 > 0.05$ within a 1-Mb window where at least one lead SNP reached a conventionally defined GWS threshold ($P < 5 \times 10^{-8}$; **Fig. 1c** and **Supplementary Fig. 1b).** An additional 19 loci reached a genome-wide marginally significant (GWM) threshold at or below $P_{META} < 1 \times 10^{-6}$, of which 12 mapped to previously reported autoimmune loci and 7 mapped to putatively novel autoimmune loci (**Fig. 1** and **Supplementary Table 2a**).

We identified five putatively novel GWS loci: *CD40LG* ($P_{META} < 8.38 \times 10^{-11}$), *ADGRL2* ($P_{META} < 8.38 \times 10^{-11}$), *TENM3* ($P_{META} < 8.38 \times 10^{-11}$), *ANKRD30A* ($P_{META} < 8.38 \times 10^{-11}$) and *ADCY7* ($P_{META} < 5.99 \times 10^{-9}$). For each lead association locus, we identified the corresponding combination of pAIDs contributing to the association signal by enumerating all 1,023 unique disease combinations (for example, one disease, T1D; two diseases, T1D and SLE; or four diseases, UC, CD, CEL and SLE) and performing association testing to identify the disease combination that yielded the maximum logistic regression *Z*-score (Online Methods)[26]. With the exception of *ANKRD30A*, the loci were jointly associated with at least two or more pAIDs; for example, *CD40LG* was shared by CEL, CD and UC (**Fig. 1** and **Table 1**). Among the 27 GWS lead SNPs, 22 had been reported previously as GWS for at least one of the associated pAIDs (specifically, for the corresponding adult phenotypes) identified by our analysis (**Supplementary Table 1b**)[12,27]. The most widely shared locus, chr4q27:rs62324212, mapping to an intronic SNP in *IL21-AS1* and residing just upstream of *IL21*, was shared across all ten diseases, and three of these associations were novel (THY, ankylosing spondylitis ~~(AS)~~ and common variable immunodeficiency (CVID)). Among the previously known GWS loci in adult-onset or generalized autoimmune disease, we identified at least one previously unrecognized pAID association for more than 50% of them (**Supplementary Table 2c**,**d**).

A number of the pAIDs were significantly associated with disease-specific signals mapping to or near the locus encoding HLA-DRB1. However, even the two most significant LD-independent variants, associated with T1D and JIA, respectively, were disease specific (**Supplementary Fig. 3**), which suggests that the variants associated with a given disease are distinct. Although some of these associated signals were shared by at least one other autoimmune disease, in no instance was a single signal associated with any of the diseases shared across all other diseases, which further underscores the complexity of signal sharing across the major histocompatibility complex (MHC) (**Supplementary Fig. 3b**).

Disease-specific and cross-autoimmune replication support for pAID-associated loci

We performed *in silico* analysis to test whether the reported associations could be replicated in an independent data set. We observed nominally significant replication support for four of the five putatively novel GWS loci, including three instances of disease-specific replication (**Supplementary Table 1d**). Among the replicated loci, chrXq26.3 (rs2807264), mapping within 70 Kb upstream of *CD40LG*, was notable, as we observed disease-specific replication in both UC ($P < 4.66 \times 10^{-5}$) and CD ($P < 5.81 \times 10^{-4}$), as well as cross-autoimmune replication in AS ($P < 9.54 \times 10^{-3}$). Although rs2807264 was not identified in our analysis as associated with pediatric AS, it is well documented that adult-onset AS and pediatric AS may be biologically different diseases with independent genetic etiologies[28,29]. A third disease-specific replication ($P < 5.99 \times 10^{-6}$) was identified in CD for the chr16q12.1 (rs77150043) signal mapping to an intronic position in *ADCY7*. This third instance and the replication of the *CD40LG* locus in UC were both significant, even after a very conservative Bonferonni adjustment for 156 tests ($P < 3.21 \times 10^{-4}$). A nominally significant pan-autoimmune replication signal ($P < 1.69 \times 10^{-2}$) was also observed at chr1p31.1 (rs2066363) near *LPHN2* in UC, and a replication signal ($P < 3.65 \times 10^{-3}$) was also observed at the chr4q35.1 locus (rs77150043) in psoriasis (**Supplementary Tables 1d** and **2e**).

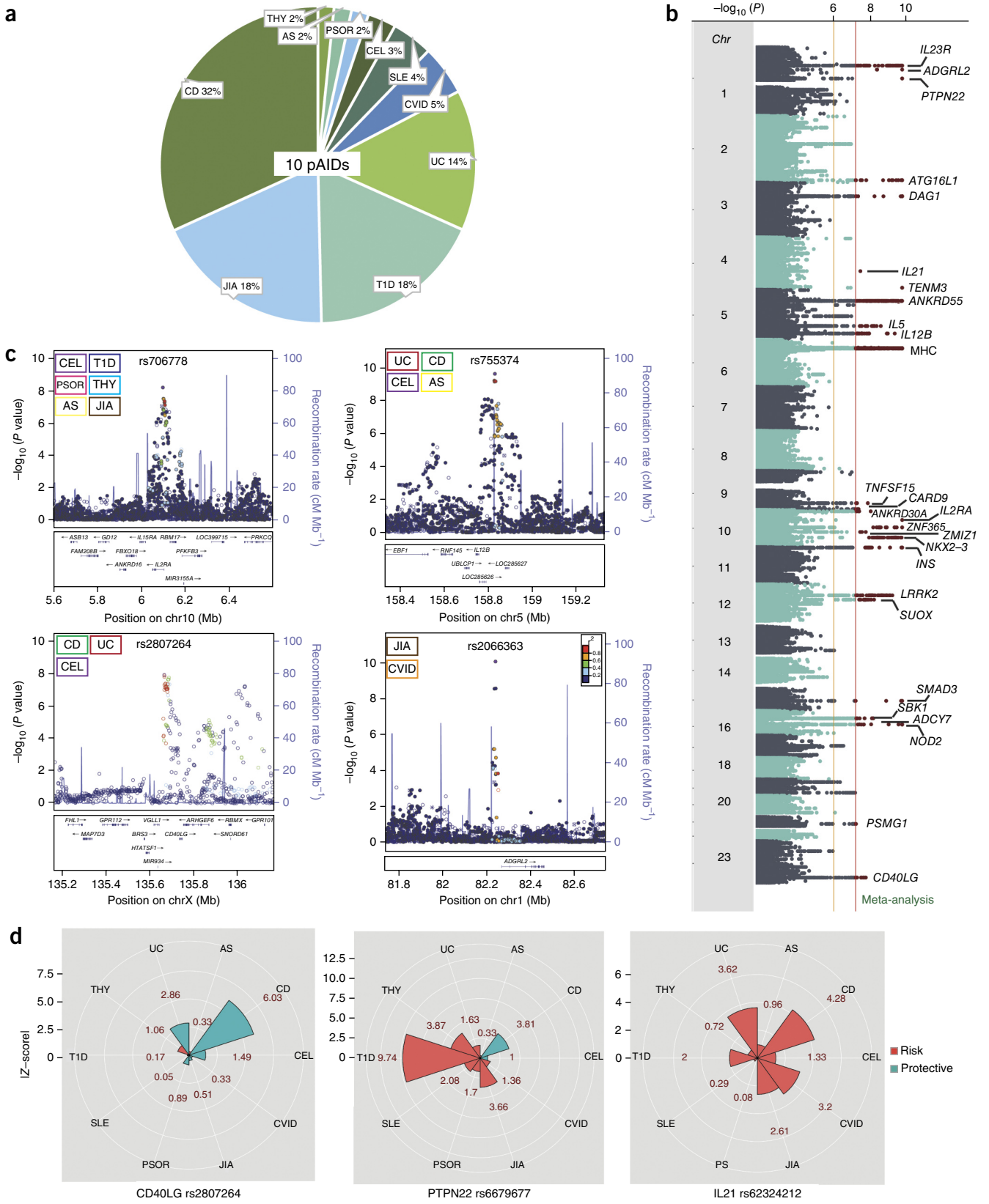
Sharing of pAID-associated SNPs and bidirectional effects of some SNPs on disease-specific risk

Of the 27 GWS loci, 81% (22) showed evidence of being shared among multiple pAIDs. These mapped to 77 different SNP-pAID combinations, 44 of which had been previously reported at or near genome-wide significance ($P < 1 \times 10^{-6}$), whereas 33 represented potentially novel disease-association signals (**Table 1** and **Supplementary Table 1**). Although *PTPN22* c.1858C>T (rs2476601) increases the risk for T1D, the variant is protective against CD[17,30–32]. We identified eight other instances ($P < 0.05$) where the risk allele shared by the model pAID combination was associated with protection against another pAID (**Fig. ~~2a~~** and **Supplementary Fig. ~~7~~**).

Figure 1 The ten pAID case cohorts and top pAID-association loci identified. (**a**) The ten pediatric autoimmune diseases studied. (**b**) Top pAID-association signals identified by inverse χ^2 meta-analysis. The top 27 loci (where at least one lead SNP reached genome-wide significance: $P_{META} < 5 \times 10^{-8}$) are annotated with the candidate gene symbol. (**c**) Novel and established pAID-association loci. Top left: rs706778 (chr10p15.1) is a known DNase I peak and an intronic SNP in *IL2RA* and was associated with THY, AS, PSOR, CEL, T1D and JIA. Top right: rs755374 (chr5q33.3) is an intergenic SNP upstream of *IL12B* and was associated with AS, CEL, UC and CD. Bottom left: rs2807264 (chrXq26.3), mapping near *CD40LG*, was associated with CEL, UC and CD, and chr15q22.33 (rs72743477), also mapping to an intronic position in *SMAD3*, was associated with UC, CD and AS. Bottom right: SNPs are colored according to pairwise LD (r^2) with respect to the most strongly associated lead SNP in the locus. Associated pAIDs are indicated at the upper left. pAID associations are color-coded according to the key in each plot. (**d**) Pleiotropic candidate genes have pleiotropic effect sizes and directions across pAIDs. Although a few pleiotropic SNPs had consistent effect directions across diseases (e.g., *IL21*), for many loci (e.g., *PTPN22* and *CLEC16A*), the candidate SNP had variable effect directions across diseases. The radii of the wedges correspond to the absolute values of the *Z*-scores (beta/s.e.) for each pAID, and the color indicates whether the SNP is protective (green) or risk-associated (red) for each disease.

Biological support of associated loci from the public domain

To integrate our results with experimental and predictive biological data, we curated four categories of SNP annotations: (1) functional: variants that are exonic, affect transcription, are miRNA targets or tag copy-number polymorphic regions; (2) regulatory: transcription factor (TF)-binding sites and DNase-hypersensitivity sites or expression quantitative trait locus (eQTL) SNPs; (3) conserved: variants with evolutionarily constrained positions or CpG islands; or (4) prior literature



Q28 **Table 1** Twenty-seven independent loci reaching genome-wide significance ($P_{META} < 5 \times 10^{-8}$) after adjustment for the use of shared controls using an inverse χ^2 meta-analysis across the pAIDs

Chr	Pos (Mb)	SNP	Region	Gene	A1	MAF	P_{META}	Known P*	pAIDs
1	67.7	rs11580078	1p31.3	*IL23R*	G	0.43	8.4×10^{-11}	1.0×10^{-146}	CD#
1	**82.2**	**rs2066363**	**1p31.1**	***ADGRL2***	**C**	**0.34**	**8.4×10^{-11}**	**Novel**	**CVID, JIA**
1	114.3	rs6679677	1p13.2	*PTPN22*	A	0.09	8.4×10^{-11}	1.1×10^{-88}	THY#, PSOR, T1D#, JIA#
2	234.2	rs36001488	2q37.1	*ATG16L1*	C	0.48	8.4×10^{-11}	1.0×10^{-12}	PSOR, CD#
3	49.6	rs4625	3p21.31	*DAG1*	G	0.31	8.4×10^{-11}	1.0×10^{-47}	PSOR#, CEL, UC#, CD#
4	123.6	rs62324212	4q27	*IL21*	A	0.42	2.6×10^{-8}	1.0×10^{-9}	THY, AS, CEL#, CVID, UC#, T1D#, JIA#, CD#
4	**183.7**	**rs7660520**	**4q35.1**	***TENM3***	**A**	**0.26**	**8.4×10^{-11}**	**Novel**	**THY, AS, CEL, SLE, CVID, JIA**
5	40.5	rs7725052	5p13.1	*PTGER4*	C	0.43	8.4×10^{-11}	1.4×10^{-10}	CD#
5	55.4	rs7731626	5q11.2	*ANKRD55*	A	0.39	1.4×10^{-10}	2.7×10^{-11}	JIA#, CD#
5	131.8	rs11741255	5q31.1	*IL5*	A	0.42	1.6×10^{-9}	1.4×10^{-52}	PSOR#, CEL, CD#
5	158.8	rs755374	5q33.3	*IL12B*	T	0.32	2.3×10^{-10}	1.4×10^{-42}	AS#, CEL, UC#, CD#
9	117.6	rs4246905	9q32	*TNFSF15*	T	0.28	9.5×10^{-9}	1.2×10^{-17}	UC#, CD#
9	139.3	rs11145763	9q34.3	*CARD9*	C	0.40	3.3×10^{-8}	1.0×10^{-6}	AS#, UC#, CD#
10	6.1	rs706778	10p15.1	*IL2RA*	T	0.41	6.3×10^{-9}	1.7×10^{-12}	THY, AS, PSOR#, CEL, T1D#, JIA#
10	**37.6**	**rs7100025**	**10p11.21**	***ANKRD30A***	**G**	**0.34**	**8.4×10^{-11}**	**Novel**	**JIA**
10	64.4	rs10822050	10q21.2	*ZNF365*	C	0.39	8.4×10^{-11}	5.0×10^{-17}	SLE, CD#
10	81.0	rs1250563	10q22.3	*ZMIZ1*	C	0.29	1.3×10^{-8}	1.1×10^{-30}	PSOR#, CD#
10	101.3	rs1332099	10q24.2	*NKX2-3*	T	0.46	9.1×10^{-11}	1.0×10^{-54}	UC#, CD#
11	2.2	rs17885785	11p15.5	*INS*	T	0.20	8.4×10^{-11}	4.4×10^{-48}	T1D#
12	40.8	rs17466626	12q12	*LRRK2*	G	0.02	3.2×10^{-10}	3.0×10^{-10}	AS, CD#
12	56.4	rs1689510	12q13.2	*SUOX*	C	0.31	4.0×10^{-9}	1.1×10^{-10}	PSOR#, T1D#
15	67.5	rs72743477	15q22.33	*SMAD3*	G	0.21	8.4×10^{-11}	2.7×10^{-19}	AS, UC, CD#
16	28.3	rs12598357	16p11.2	*SBK1*	G	0.39	4.4×10^{-9}	1.0×10^{-8}	THY, AS#, PSOR, CEL, UC, CD#
16	**50.3**	**rs77150043**	**16q12.1**	***ADCY7***	**T**	**0.23**	**6.0×10^{-9}**	**Novel**	**PSOR, CD**
16	50.7	rs117372389	16q12.1	*NOD2*	T	0.02	8.4×10^{-11}	2.9×10^{-69}	CD#
21	40.5	rs2836882	21q22.2	*PSMG1*	A	0.27	4.8×10^{-8}	2.8×10^{-14}	UC#, CD#
23	**135.7**	**rs2807264**	**Xq26.3**	***CD40LG***	**C**	**0.21**	**1.3×10^{-8}**	**Novel**	**CEL, UC, CD**

Chr, chromosome; Pos (Mb), position in *hg19*; Region, cytogenetic band; A1, alternative allele; MAF, minor allele frequency (controls); Known P*, lowest P value from published association studies. Pound symbols (#) denote previously reported disease-associated SNPs. "Novel" denotes new loci (bolded) that reached genome-wide significance for the first time in the present study (to our knowledge).

support: a gene or locus previously reported to be associated with autoimmune diseases or immune function. Indeed, 100% of the GWS lead SNPs or their nearby LD proxies ($r^2 > 0.8$ on the basis of 1KGP-RP within 500 Kb up- or downstream) belonged to one or more of these categories (**Fig. 3a**). Nevertheless, the majority of the 27 GWS SNPs did not confer transcriptional consequences (51% were intronic variants

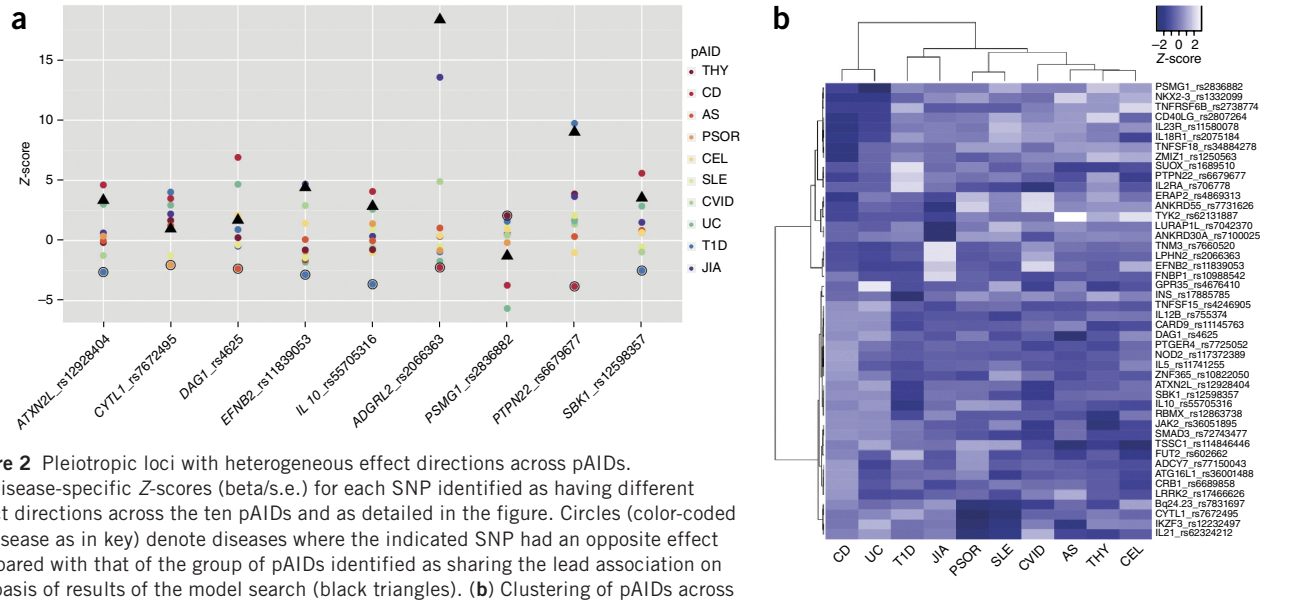


Figure 2 Pleiotropic loci with heterogeneous effect directions across pAIDs. (**a**) Disease-specific *Z*-scores (beta/s.e.) for each SNP identified as having different effect directions across the ten pAIDs and as detailed in the figure. Circles (color-coded by disease as in key) denote diseases where the indicated SNP had an opposite effect compared with that of the group of pAIDs identified as sharing the lead association on the basis of results of the model search (black triangles). (**b**) Clustering of pAIDs across the lead loci on the basis of disease-specific effect sizes. Agglomerative hierarchical clustering across ten pAIDs on the basis of normalized directional *Z*-scores (beta/s.e.) resulting from logistic regression analysis in each disease for the 27 lead loci based on those disease combinations identified by the model search analysis as producing the strongest association-test statistics.

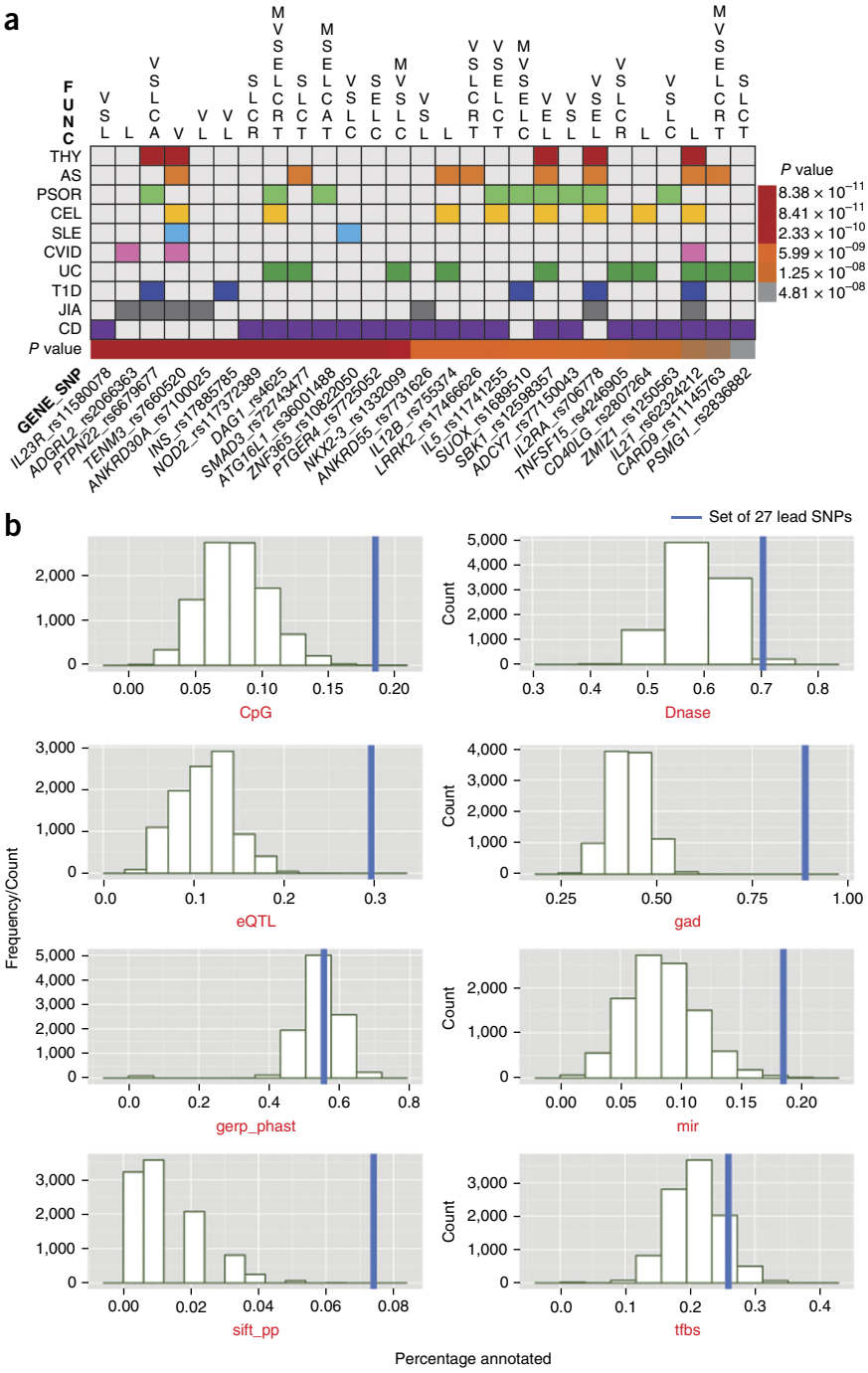


Figure 3 Integrated annotation of pAID-association loci using existing predictive and experimental data sets. (**a**) Biological, functional and literature annotations for the 27 loci reaching genome-wide significance in meta-analysis. Loci (identified by the lead SNPs and candidate genes) are organized by column; the colors in the table denote the associated pAIDs, functional annotations are presented at the top of the table, and the color bar at the bottom represents the meta-analysis P_{meta} values (according to key at right). For each locus, the lead SNP and proxy SNPs ($r^2 > 0.8$) were included in the annotation protocol (Online Methods). (**b**) Distribution and enrichment of experimental and predicted annotations for the top 27 GWS SNPs. The annotation frequencies were used to calculate the relative enrichment of pAID SNPs (blue bars) as compared with that of 10,000 random 100-SNP sets drawn from the genome in each annotation category. CpG, CpG islands; DNase, DNase-hypersensitivity I sites; gad, known genetic association; gerp_phast, conserved positions; mir, miRNAs; sift_pp, functional mutations in SIFT; tfbs, TF–binding sites.

and 28% were intergenic or up- or downstream gene variants), which suggests that many of these SNPs either tag the true causal variants or affect disease risk through regulatory and/or epigenetic mechanisms (**Fig. 3b**).

To determine whether the set of pAID-associated SNPs was enriched for specific annotation categories, we compared its annotation percentage with the percentages of 10,000 simulated sets of SNPs with MAF > 0.01 drawn from 1KGP-RP for each category. We found that pAID-associated SNPs were enriched for CpG islands ($P_{perm} < 1.0 \times 10^{-4}$), TF-binding sites ($P_{perm} < 3.4 \times 10^{-3}$) and miRNA-binding sites ($P_{perm} < 1.0 \times 10^{-4}$), among other findings of biological disease relevance (**Supplementary Fig. 1d**,**e**).

Candidate pAID genes share expression profiles across immune cell types and tissues

Recent studies show that gene-based association testing (GBAT) may boost the power of genetic discovery[33–35]. We performed GBAT (with VEGAS[33]) using genome-wide summary-level P_{META} values. We identified 182 significant pAID-associated genes (simulation-based $P_{sim} < 2.80 \times 10^{-6}$) on the basis of a Bonferonni adjustment for ~17,500 protein-coding genes in the genome (**Supplementary Table 3a**). To illustrate the biological relevance of this set of genes, we examined their transcript levels in a human gene expression microarray data set consisting of 12,000 genes and 126 tissue and/or cell types[36]. pAID-associated gene expression across immune tissues or cell types (ES-I, 4.05) was notably higher than that across non-immune types (ES-NI, 2.10) on the basis of a one-tailed Wilcoxon rank-sum test ($P < 1.66 \times 10^{-10}$). When all extended MHC genes were excluded, the average expression of pAID-associated genes remained significantly higher ($P < 1.27 \times 10^{-7}$) for immune (1.043) than for non-immune (0.648) tissues and cell types. The immune-specific enrichment of pAID-associated gene transcripts was comparable to that observed in adult cohorts[12]; comparatively, schizophrenia-associated genes showed no such enrichment (**Fig. 4a**). We observed similar results when we used the Kolmogorov-Smirnov test (**Supplementary Fig. 5a**).

We examined the expression of pAID genes across a whole-transcriptome data set comprising more than 200 murine immune cell types isolated by flow cytometry (ImmGen[37]; Online Methods and **Supplementary Table 3c**). Genes associated with pAIDs demonstrated differential expression across immune cell types (**Supplementary Fig. 5b**) and showed higher expression than genes associated with non-immune traits, similar to results observed from human tissue

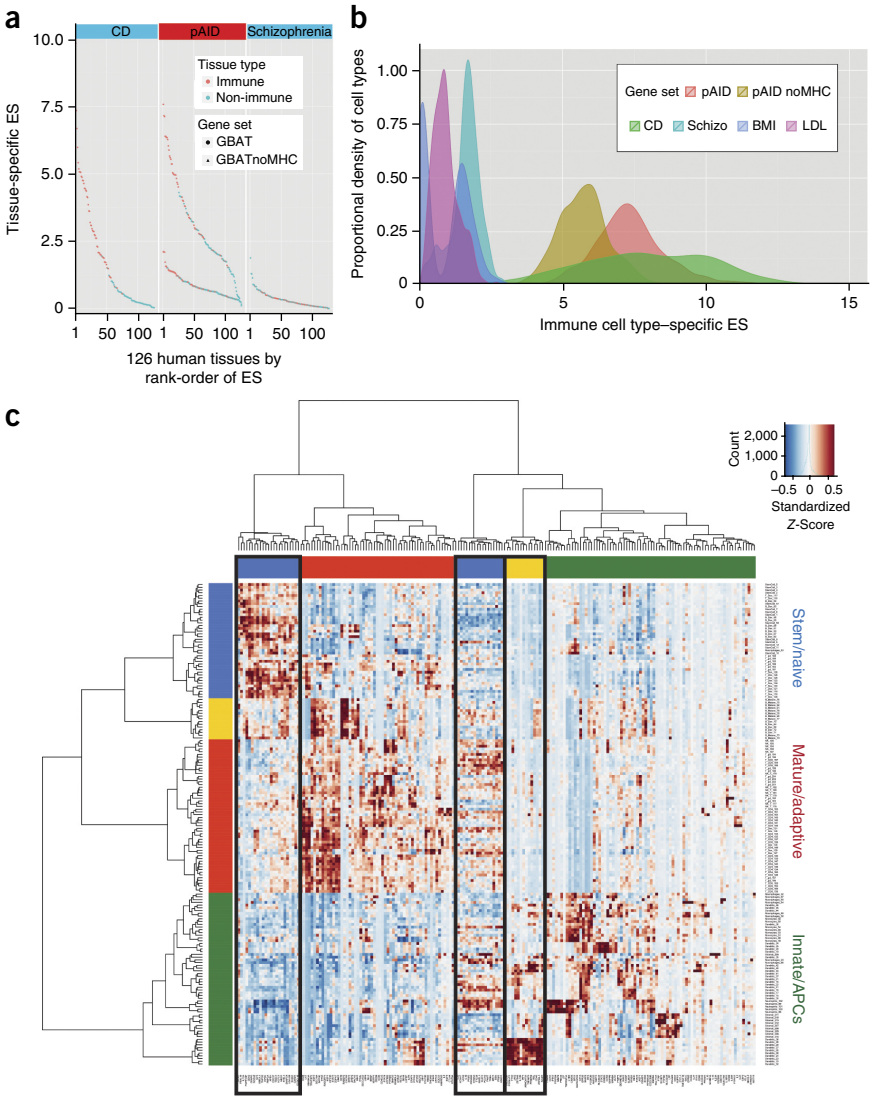


Figure 4 Tissue-specific gene set enrichment analysis (TGSEA) of pediatric and adult autoimmune data sets identifies autoimmune-associated gene expression patterns across immune cells and tissues. (**a**) Expression enrichment of autoimmune-associated genes across human tissues. Distribution of TGSEA enrichment score (ES) values across 126 tissues for pAID-associated genes (center) either with (circles) or without (triangles) the extended MHC ~~(for clarity, also labeled in the plot with + or −, respectively)~~. Results for the pAID gene set are compared with those obtained for known genes associated with CD (left) and schizophrenia (right). Tissue and cell types are classified as immune (red) or non-immune (blue) and are ranked left to right on the basis of the magnitude of the ES test statistic. (**b**) Enrichment of pAID-associated gene expression across diverse murine immune cell types. Distribution of pAID-associated gene ES values across murine immune cell types either including (red) or excluding the genes within the MHC (yellow); results are compared with those for genes associated with CD, schizophrenia (Schizo, turquoise), LDL cholesterol (LDL, magenta) or body mass index (BMI, blue) abstracted from the National Human Genome Research Institute (NHGRI) GWAS Catalog. (**c**) Hierarchical clustering based on the expression of pleiotropic candidate genes associated with three or more autoimmune diseases across the murine immune cells. Boxes outlined in black denote gene clusters enriched for specific disease associations discussed in the text.

data (**Fig. 4b**). As the expression levels of these 'pleiotropic' genes varied diversely across immune cell types, we performed agglomerative hierarchical clustering to identify sets of genes sharing similar profiles. Genes that belonged to the same cluster (and thus shared similar expression profiles) were found to be enriched for association with specific individual or multiple autoimmune diseases (**Fig. 4c**). For example, cluster 1 genes, such as *ICAM1*, *CD40*, *JAK2*, *TYK2* and *IL12B*, with known roles in immune effector cell activation and proliferation, were enriched for association with PSC and UC and were associated with both diseases ($P < 6.82 \times 10^{-4}$, one-tailed Fisher's exact test), and the expression of these genes was highest in a small subset of CD11b^{+} lymphoid dendritic cells. These findings are consistent with the clinical observation that as many as 80% of patients diagnosed with PSC have been diagnosed with UC, and that the risk of PSC is approximately 600-fold higher in patients with UC[38,39]. Cluster 2 genes included genes encoding a number of cytokines and cytokine-response factors, such as *IL19*, *IL20*, *STAT5A* and *IL2RA*, the products of which regulate effector T cell activation, differentiation and proliferation. All of these were more broadly expressed across mature natural killer (NK) cells, NK T cells and T cells, as well as neutrophils. This cluster of genes was enriched for association with MS ($P < 9.8 \times 10^{-4}$), with CEL (marginally) ($P < 0.062$) and with both diseases ($P < 3.41 \times 10^{-4}$). Genes encoding nucleic acid–binding proteins, such as *ILF3*, *CENPO*, *MED1* and *NCOA3*, were enriched in cluster 3. Genes in this cluster were jointly associated with SLE and PS ($P < 0.03$), which is consistent with experimental and clinical data demonstrating that early defects in B cell[40,41] and T cell[42–44] clonal selection, respectively, may have important roles in the etiology of these diseases.

Quantification of genetic risk factors shared across pAIDs

We developed a novel method to specifically examine genome-wide pairwise-association signal sharing (referred to as a GPS test) across the pAIDs (Online Methods). Only data from the genotyped pAID cohort were used for this analysis. After Bonferroni adjustment for 45 pairwise combinations, the GPS test identified evidence of sharing between a number of pAID pairs noted in prior reports on autoimmune disease, including T1D-CEL ($P_{gps} < 3.44 \times 10^{-5}$), T1D-THY ($P_{gps} < 2.03 \times 10^{-3}$) UC-CD ($P_{gps} < 2.36 \times 10^{-3}$) and AS-PS ($P_{gps} < 8.15 \times 10^{-3}$). We also identified a strong GPS score for JIA-CVID ($P_{gps} < 6.88 \times 10^{-5}$). The correlations between JIA-CVID ($P_{gps} < 7.30 \times 10^{-5}$) and UC-CD ($P_{gps} < 7.32 \times 10^{-4}$) were more significant after the exclusion of markers from within the MHC region (**Supplementary Fig. ~~7~~b**).

Finally, we examined evidence of sharing across the full range of autoimmune diseases using ImmunoBase[27]. We identified significant associations between UC-CD ($P < 2.15 \times 10^{-4}$) and JIA-CVID ($P < 1.44 \times 10^{-6}$), along with a number of novel pairwise relationships that included autoimmune diseases other than the ten in this study, such as that between ~~SJO-SS~~ ($P < 1.30 \times 10^{-28}$) and PBC-SJO Q10

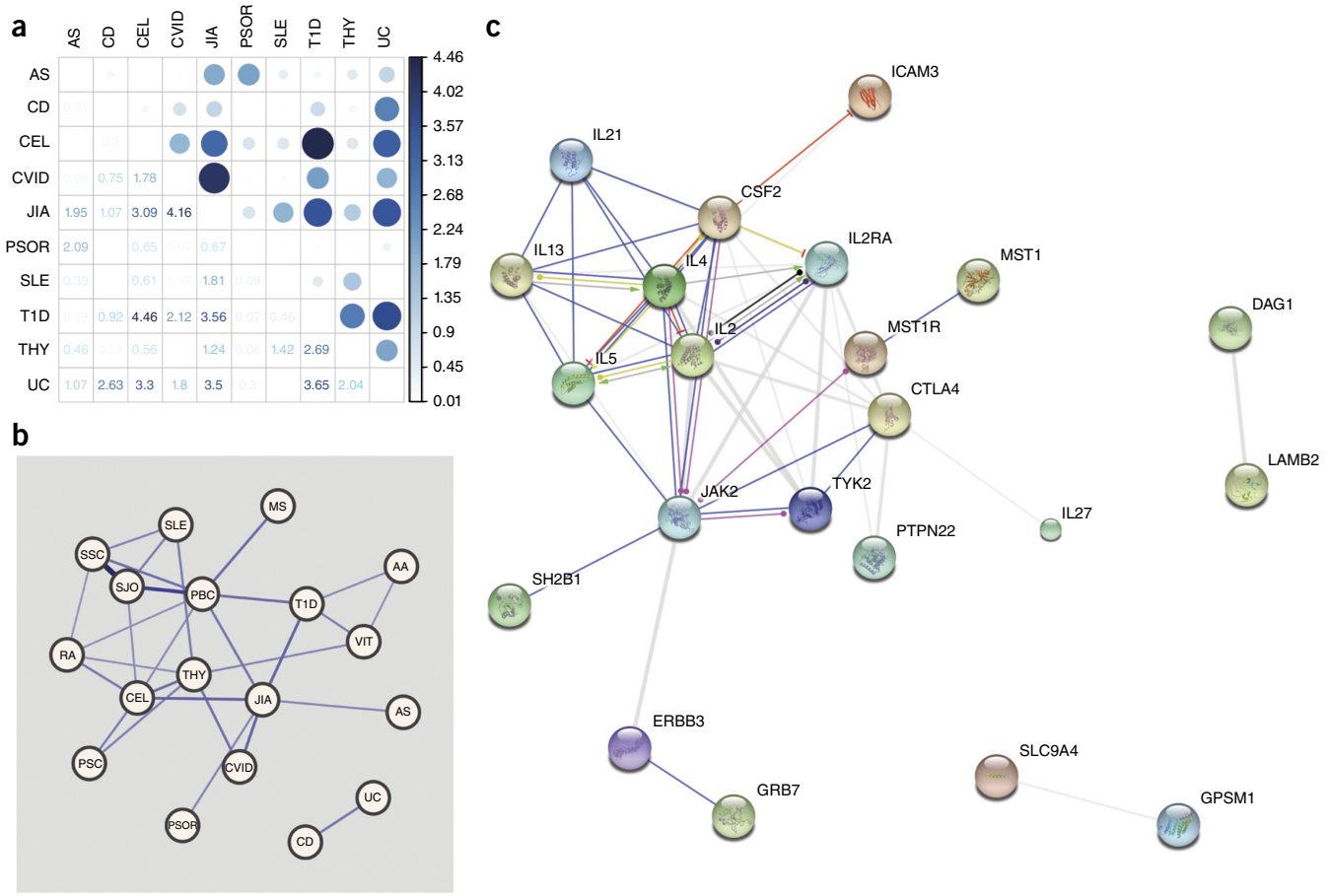


Figure 5 Genetic variants shared across the ten pAIDs reveal autoimmune disease networks. (**a**) Quantification of pAID genetic sharing by GPS test including SNPs within the extended MHC. Correlation plot of results of the pairwise pAID GPS test; the color intensity and the size of each circle are proportional to the strength of the correlation as the negative base ten logarithms of the GPS test *P* values (color-coded numbers in squares). (**b**) Quantification of autoimmune disease genetic sharing by locus-specific pairwise sharing. Undirected weighted network graph depicting results from the LPS test. Edge size represents the magnitude of the LPS test statistic; labeled nodes for each of the 17 autoimmune diseases are positioned on the basis of a force-directed layout. Edges represent significant pairs after Bonferroni adjustment ($P_{adj} < 0.05$). (**c**) Protein-protein interaction network analysis of the top pAID-associated protein candidates in STRING; action view of protein interactions observed across the top 46 GWM ($P < 1 \times 10^{-6}$) signals, of which 44 could be mapped to corresponding proteins. Views were generated on the basis of results for known and predicted protein interactions produced by the STRING DB *Homo sapiens* database. The plots shown are results of the 'action' view, where the molecular actions (stimulatory, repressive or binding) are illustrated by arrows.

($P < 3.86 \times 10^{-12}$). We plotted those relationships that were significant after Bonferroni adjustment for 153 pairwise tests using an undirected weighted network (**Fig. 5b** and **Supplementary Table 4**). Collectively, these results support genetic sharing between the various autoimmune diseases and allow for further refinement of the shared signals, potentially enabling the application of targeted therapeutic interventions at multiple levels, such as along the CD40L-CD40, JAK-STAT and T_H1/T_H2-T_H17-interleukin signaling pathways.

DISCUSSION

A major goal of this study was to identify shared genetic etiologies across pAIDs and illustrate how they jointly and disparately affect pAID susceptibility. Knowledge of shared genetic etiologies may help pinpoint common therapeutic mechanisms, especially since certain pAIDs (for example, THY, CEL and T1D) exhibit high rates of comorbidity and concordance in twins with others (for example, CD and UC) being clustered in families[9,19,45,46].

Of the 27 GWS pAID-association loci identified, 81% were shared by at least two pAIDs (**Table 1** and **Supplementary Table 1**). Moreover, 5 of the 27 loci were novel signals not previously reported at GWS levels in association with autoimmune diseases, including chr1p31.1 (rs2066363), mapping near *ADGRL2*, a gene that encodes a member of the latrophilin subfamily of G protein–coupled receptors that regulates exocytosis. Although this signal was associated with JIA and CVID, a microsatellite study of PBC in a Japanese cohort localized an association signal to a 100-Kb region enclosing *ADGRL2* (ref. 47). Nominally significant replication support at this locus was identified in the adult UC cohort from the ~~IBD Consortium~~ (REF). Both JIA and CVID are among the six pAIDs (THY, AS, CEL, SLE, CVID and JIA) associated with the chr4q35.1 locus (rs7660520), which resides just downstream of *TENM3*. The observed association with a broad range of pAIDs may be related to eQTL signals in *TENM3* SNPs that correlate with serum eosinophil counts[48] and immunoglobulin G (IgG) glycosylation rates; the latter was referenced in a study showing a pleiotropic role for IgG glycosylation–associated SNPs in autoimmune-disease risk susceptibility[49]. The third novel association was identified near chr10p11.21 (rs7100025), mapping to TF gene *ANKRD30A*, which encodes an antigen recognized by $CD8^+$ T cell clones[50]. The fourth signal was

associated with the inflammatory diseases PS and CD near chr16q12.1 (rs77150043). This intronic SNP in *ADCY7* encodes a member of the adenylate cyclase enzyme family and is strongly expressed in peripheral leukocytes, spleen, thymus and lung tissues[51], and it is supported by data from studies in mice[52]. The fifth novel signal, rs34030418, mapping near *CD40LG* and associated with CEL, UC and CD, is the ligand of the prominent TNF superfamily receptor CD40 (refs. 53,54). The CD40 ligand is a particularly compelling candidate, as the locus encoding the CD40 receptor is an established GWAS locus in RA and MS, has been functionally studied in cell culture and animal models, and was the focus of a recent large-scale RA drug-screening effort[55].

A set of GWS candidate SNPs were enriched for miRNA and TF-binding sites. We performed a gene-set enrichment analysis[56] using GBAT and identified 39 significant ($P_{BH} < 0.05$) miRNAs, including as top candidates two well-known miRNA families, miR-22 and miR-135a (**Supplementary Table 5a**). miR-135a has been shown to target IRS2, a regulator of insulin signaling and glucose uptake, in model systems[57]. Our candidate genes were enriched for targets of dozens of TFs, with the most prominent being SP1 ($P_{BH} < 2.30 \times 10^{-12}$), NFAT ($P_{BH} < 8.54 \times 10^{-9}$) and NFKB ($P_{BH} < 1.03 \times 10^{-8}$) (**Supplementary Table 5b**).

Using GBAT with DAVID[58], GSEA[36], IPA[59] and Pathway Commons[60], among others, we identified strong enrichment for proteins that act in cytokine signaling; antigen processing and presentation; T cell activation; JAK-STAT activation; and T_H1-, T_H2- and T_H17-associated cytokine signaling (**Supplementary Table 6**). Of these pathways, JAK2 signaling was particularly compelling ($P_{BH} < 6.93 \times 10^{-5}$; ~~**Supplementary Fig. 6b**~~), consistent with the enrichment of known protein-protein interactions ($P_{STRING} < 1 \times 10^{-20}$) (**Supplementary Fig. 6**). We also uncovered evidence supporting shared genetic susceptibility for disease pairs that have not yet been well established (for example, JIA-CVID). The association between JIA and CVID is noteworthy, given that CVID actually represents a group of complex immunodeficiencies rather than a classic autoimmune disease. When we examined the overlap between CVID and ~~all~~ other pAIDs using both GPS ($P_{adj} < 3.10 \times 10^{-3}$) and locus-specific pairwise sharing (LPS) ($P_{adj} < 1.47 \times 10^{-8}$) network analysis tests, we consistently observed overrepresentation of interaction between CVID and JIA (**Fig. 5** and **Supplementary Fig. ~~7b~~**). Our results show that more than 70% (19) of the 27 GWS loci we identified were shared by at least three autoimmune diseases (**Table 1**), including both previously reported (for example, *IL2RA*[6] and *IL12B*[4]) and novel (for example, *TENM3* ~~(ref. 6)~~ and *CD40LG*[3]) signals. Moreover, using tissue-specific gene set enrichment analysis, we not only highlighted the expected enrichment of genes associated with CEL and SLE in γδ T cells, $CD4^+$ T cells and NK T cells but also identified interesting joint enrichment of genes associated with PSC and UC in a set of mature dendritic cells (**Fig. 4c**).

Many of the shared risk ~~factors~~ in pAIDs affect genes encoding proteins that are established therapeutic targets (for example, CD40L and CD40 (refs. 54,55)), and a number of the genes identified here have diverse biological effects and are currently being explored for clinical uses. Consequently, drug-repurposing approaches may present feasible options in pAIDs, where these gene networks and pathways could be targeted in an expedited manner.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the subjects and their families for their participation in genotyping studies and the Biobank Repository at the Center for Applied Genomics. We acknowledge M.V. Holmes, H. Matsunami, L. Steel and E. Carrigan for their technical assistance and review of the manuscript. We are also thankful for the contributions of the Italian IBD Group, including S. Cucchiara (Roma), P. Lionetti (Firenze), G. Barabino (Genova), G.L. de Angelis (Parma), G. Guariso (Padova), C. Catassi (Ancona), G. Lombardi (Pescara), A.M. Staiano (Napoli), D. De Venuto (Bari), C. Romano (Messina), R. D'incà (Padova), M. Vecchi (Milano), A. Andriulli and F. Bossa (S. Giovanni Rotondo). The data sets used for the replication analyses were obtained through dbGaP accession numbers phs000344, phs000127, phs000274, phs000171, phs000224, phs000130, phs000019, phs000091, phs000206, phs000168, phs000138, phs000125 and phs000092. We thank the NIH data repository, the contributing investigators who contributed the phenotype data and DNA samples from their original studies, and the primary funding organizations that supported the contributing investigators. This study made use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113. Y.R.L. is supported by the Paul and Daisy Soros Fellowship for New Americans and the NIH F30 Individual NRSA Training Grant. This study was supported by Institutional Development Funds from The Children's Hospital of Philadelphia and by DP3DK085708, RC1AR058606, U01HG006830, the Crohn's & Colitis Foundation of America, the Juvenile Diabetes Research Foundation, NIH grant CA127334 (to H.L. and S.D.Z.) and a grant from the Lupus Research Institute (to E.T.L.P.). This work was supported in part by the NIH (grant R01-HG006849 to A.K.). F.G. is a Howard Hughes Medical Institute International Student Research fellow.

AUTHOR CONTRIBUTIONS

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Cooper, G.S., Bynum, M.L. & Somers, E.C. Recent insights in the epidemiology of autoimmune diseases: improved prevalence estimates and understanding of clustering of diseases. *J. Autoimmun.* **33**, 197–207 (2009).
2. Cooper, J.D. *et al.* Seven newly identified loci for autoimmune thyroid disease. *Hum. Mol. Genet.* **21**, 5202–5208 (2012).
3. Tsoi, L.C. *et al.* Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat. Genet.* **44**, 1341–1348 (2012).
4. Hinks, A. *et al.* Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nat. Genet.* **45**, 664–669 (2013).
5. Liu, J.Z. *et al.* Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis. *Nat. Genet.* **44**, 1137–1141 (2012).
6. Liu, J.Z. *et al.* Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nat. Genet.* **45**, 670–675 (2013).
7. Eyre, S. *et al.* High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis. *Nat. Genet.* **44**, 1336–1340 (2012).
8. Zhernakova, A. *et al.* Meta-analysis of genome-wide association studies in celiac disease and rheumatoid arthritis identifies fourteen non-HLA shared loci. *PLoS Genet.* **7**, e1002004 (2011).
9. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
10. International Multiple Sclerosis Genetics Consortium. *et al.* Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476**, 214–219 (2011).
11. Beecham, A.H. *et al.* Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat. Genet.* **45**, 1353–1360 (2013).
12. National Human Genome Research Institute Published Genome-Wide Associations through 08/01/2014. *NHGRI ~~GWA~~ Catalog* http://www.genome.gov/multimedia/illustrations/GWAS_2011_3.pdf (2014).
13. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, D1001–D1006 (2014).
14. Cortes, A. & Brown, M.A. Promise and pitfalls of the Immunochip. *Arthritis Res. Ther.* **13**, 101 (2011).
15. Hakonarson, H. *et al.* A genome-wide association study identifies *KIAA0350* as a type 1 diabetes gene. *Nature* **448**, 591–594 (2007).
16. Hinks, A. *et al.* Association between the *PTPN22* gene and rheumatoid arthritis and juvenile idiopathic arthritis in a UK population: further support that *PTPN22* is an autoimmunity gene. *Arthritis Rheum.* **52**, 1694–1699 (2005).
17. Smyth, D.J. *et al.* Shared and distinct genetic variants in type 1 diabetes and celiac disease. *N. Engl. J. Med.* **359**, 2767–2777 (2008).
18. Harley, J.B. *et al.* Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in *ITGAM*, *PXK*, *KIAA1542* and other loci. *Nat. Genet.* **40**, 204–210 (2008).

19. Ramos, P.S. *et al.* A comprehensive analysis of shared loci between systemic lupus erythematosus (SLE) and sixteen autoimmune diseases reveals limited genetic overlap. *PLoS Genet.* **7**, e1002406 (2011).
20. Cotsapas, C. *et al.* Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet.* **7**, e1002254 (2011).
21. Cotsapas, C. & Hafler, D.A. Immune-mediated disease genetics: the shared basis of pathogenesis. *Trends Immunol.* **34**, 22–26 (2013).
22. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).
23. Delaneau, O., Coulonges, C. & Zagury, J.-F. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* **9**, 540 (2008).
24. Marchini, J. SNPTEST (v2.5) https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html (2007).
25. Zaykin, D.V. & Kozbur, D.O. *P*-value based analysis for shared controls design in genome-wide association studies. *Genet. Epidemiol.* **34**, 725–738 (2010).
26. Bhattacharjee, S. *et al.* A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am. J. Hum. Genet.* **90**, 821–835 (2012).
27. Institute for Systems Biology and Juvenile Diabetes Research Foundation–Wellcome Trust Diabetes and Inflammation Laboratory. ImmunoBase http://www.immunobase.org (2013).
28. Gensler, L.S. *et al.* Clinical, radiographic and functional differences between juvenile-onset and adult-onset ankylosing spondylitis: results from the PSOAS cohort. *Ann. Rheum. Dis.* **67**, 233–237 (2008).
29. Lin, Y.-C., Liang, T.-H., Chen, W.-S. & Lin, H.-Y. Differences between juvenile-onset ankylosing spondylitis and adult-onset ankylosing spondylitis. *J. Chin. Med. Assoc.* **72**, 573–580 (2009).
30. Anaya, J.-M., Gómez, L. & Castiblanco, J. Is there a common genetic basis for autoimmune diseases? *Clin. Dev. Immunol.* **13**, 185–195 (2006).
31. De Jager, P.L. *et al.* Evaluating the role of the 620W allele of protein tyrosine phosphatase PTPN22 in Crohn's disease and multiple sclerosis. *Eur. J. Hum. Genet.* **14**, 317–321 (2006).
32. Zhernakova, A. *et al.* Differential association of the PTPN22 coding variant with autoimmune diseases in a Dutch population. *Genes Immun.* **6**, 459–461 (2005).
33. Liu, J.Z. *et al.* A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **87**, 139–145 (2010).
34. Li, M.-X., Gui, H.-S., Kwan, J.S.H. & Sham, P.C. GATES: a rapid and powerful gene-based association test using extended Simes procedure. *Am. J. Hum. Genet.* **88**, 283–293 (2011).
35. Huang, H., Chanda, P., Alonso, A., Bader, J.S. & Arking, D.E. Gene-based tests of association. *PLoS Genet.* **7**, e1002177 (2011).
36. Benita, Y. *et al.* Gene enrichment profiles reveal T-cell development, differentiation, and lineage-specific transcription factors including ZBTB25 as a novel NF-AT repressor. *Blood* **115**, 5376–5384 (2010).
37. Heng, T.S.P. & Painter, M.W. The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.* **9**, 1091–1094 (2008).
38. Olsson, R. *et al.* Prevalence of primary sclerosing cholangitis in patients with ulcerative colitis. *Gastroenterology* **100**, 1319–1323 (1991).
39. Feld, J.J. & Heathcote, E.J. Epidemiology of autoimmune liver disease. *J. Gastroenterol. Hepatol.* **18**, 1118–1128 (2003).
40. Yurasov, S. *et al.* Defective B cell tolerance checkpoints in systemic lupus erythematosus. *J. Exp. Med.* **201**, 703–711 (2005).
41. Cappione, A. *et al.* Germinal center exclusion of autoreactive B cells is defective in human systemic lupus erythematosus. *J. Clin. Invest.* **115**, 3205–3216 (2005).
42. Evenou, J.-P. *et al.* The potent protein kinase C-selective inhibitor AEB071 (sotrastaurin) represents a new class of immunosuppressive agents affecting early T-cell activation. *J. Pharmacol. Exp. Ther.* **330**, 792–801 (2009).
43. Jegasothy, B.V. Tacrolimus (FK 506)—a new therapeutic agent for severe recalcitrant psoriasis. *Arch. Dermatol.* **128**, 781–785 (1992).
44. Nograles, K.E. & Krueger, J.G. Anti-cytokine therapies for psoriasis. *Exp. Cell Res.* **317**, 1293–1300 (2011).
45. Ergür, A.T. *et al.* Celiac disease and autoimmune thyroid disease in children with type 1 diabetes mellitus: clinical and HLA-genotyping results. *J. Clin. Res. Pediatr. Endocrinol.* **2**, 151–154 (2010).
46. Eyre, S. *et al.* Overlapping genetic susceptibility variants between three autoimmune disorders: rheumatoid arthritis, type 1 diabetes and coeliac disease. *Arthritis Res. Ther.* **12**, R175 (2010).
47. Joshita, S. *et al. A2BP1* as a novel susceptible gene for primary biliary cirrhosis in Japanese patients. *Hum. Immunol.* **71**, 520–524 (2010).
48. Pruitt, K., Brown, G., Tatusova, T. & Maglott, D. The Reference Sequence (RefSeq) database http://www.ncbi.nlm.nih.gov/books/NBK21091/ (2012).
49. Lauc, G. *et al.* Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers. *PLoS Genet.* **9**, e1003225 (2013).
50. Jäger, D. *et al.* Humoral and cellular immune responses against the breast cancer antigen NY-BR-1: definition of two HLA-A2 restricted peptide epitopes. *Cancer Immun.* **5**, 11 (2005).
51. Ludwig, M.-G. & Seuwen, K. Characterization of the human adenylyl cyclase gene family: cDNA, gene structure, and tissue distribution of the nine isoforms. *J. Recept. Signal Transduct. Res.* **22**, 79–110 (2002).
52. Jiang, L.I., Sternweis, P.C. & Wang, J.E. Zymosan activates protein kinase A via adenylyl cyclase VII to modulate innate immune responses during inflammation. *Mol. Immunol.* **54**, 14–22 (2013).
53. Anderson, D.M. *et al.* A homologue of the TNF receptor and its ligand enhance T-cell growth and dendritic-cell function. *Nature* **390**, 175–179 (1997).
54. Miyashita, T. *et al.* Bidirectional regulation of human B cell responses by CD40–CD40 ligand interactions. *J. Immunol.* **158**, 4620–4633 (1997).
55. Li, G. *et al.* Human genetics in rheumatoid arthritis guides a high-throughput drug screen of the CD40 signaling pathway. *PLoS Genet.* **9**, e1003487 (2013).
56. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res.* **41**, W77–W83 (2013).
57. Agarwal, P., Srivastava, R., Srivastava, A.K., Ali, S. & Datta, M. miR-135a targets IRS2 and regulates insulin signaling and glucose uptake in the diabetic gastrocnemius skeletal muscle. *Biochim. Biophys. Acta* **1832**, 1294–1303 (2013).
58. Huang, D.W. *et al.* The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* **8**, R183 (2007).
59. Ingenuity Systems. *Ingenuity Pathway Analysis* http://www.ingenuity.com/products/ipa (2015).
60. Cerami, E.G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39**, D685–D690 (2011).

[1]The Center for Applied Genomics, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. [2]Medical Scientist Training Program, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [3]Department of Biostatistics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [4]Division of Allergy and Immunology, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. [5]Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York, USA. [6]Program in Computational Biology and Medicine, Cornell University, Ithaca, New York, USA. [7]Department of Computer Science, New Jersey Institute of Technology, Newark, New Jersey, USA. [8]Department of Biochemistry and Molecular Biology, Eberly College of Science, The Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, Pennsylvania, USA. [9]Department of Rheumatology, Oslo University Hospital, Rikshospitalet, Oslo, Norway. [10]Division of Gastroenterology, The Center for Inflammatory Bowel Disease, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA. [11]Divison of Rheumatology, Cincinnati Children's Hospital Medical Center, Cincinnati, Ohio, USA. [12]Division of Rheumatology, Children's Mercy Hospitals and Clinics, Kansas City, Missouri, USA. [13]Department of Pediatrics, University of Utah School of Medicine and Primary Children's Medical Center, Salt Lake City, Utah, USA. [14]Division of Gastroenterology, IRCCS Casa Sollievo della Sofferenza, San Giovanni Rotondo, Italy. [15]Division of Pediatric Allergy and Immunology, University of Miami Miller School of Medicine, Miami, Florida, USA. [16]Institute of Immunology and Department of Medicine, Mount Sinai School of Medicine, New York, New York, USA. [17]Department of Paediatric Gastroenterology, Yorkhill Hospital for Sick Children, Glasgow, Scotland, UK. [18]Paediatric Gastroenterology and Nutrition, Royal Hospital for Sick Children, University of Edinburgh, Ediburgh, UK. [19]Mount Sinai Hospital IBD Centre, University of Toronto, Toronto, Ontario, Canada. [20]Unit of Gastroenterology, Department of Medical and Surgical Specialties, Careggi University Hospital, Florence, Italy. [21]Department of Immunology, Oslo University Hospital, Rikshospitalet, Oslo, Norway. [22]Department of Rheumatology, Texas Scottish Rite Hospital for Children, Dallas, Texas, USA. [23]Department of Pediatrics, Pediatric IBD Center, Cedars Sinai Medical Center, Los Angeles, California, USA. [24]Department of Pathology, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. [25]Department of Pediatrics, The Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [26]Department of Translational Medical Science, Section of Pediatrics, University of Naples Federico II, Naples, Italy. [27]Department of Pediatrics, Emory University School of Medicine and Children's Health Care of Atlanta, Atlanta, Georgia, USA. [28]Genes, Environment and Complex Disease, Murdoch Children's Research Institute, Parkville, Victoria, Australia. [29]Department of Pediatrics, University of Melbourne, Parkville, Victoria, Australia. [30]Pediatric Rheumatology Unit, Royal Children's Hospital, Parkville, Victoria, Australia. [31]Arthritis and Rheumatology Research, Murdoch Children's Research Institute, Parkville, Victoria, Australia. [32]Sarah M. and Charles E. Seay Center for Musculoskeletal Research, Texas Scottish Rite Hospital for Children, Dallas, Texas, USA. [33]Department of Clinical Immunology, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [34]Section of Immunology, Allergy, and Rheumatology, Department of Pediatric Medicine, Texas Children's Hospital, Houston, Texas, USA. [35]Division of Rheumatology, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. [36]The Hospital for Sick Children, University of Toronto, Toronto, Ontario, Canada. [37]Gastrointestinal Unit, Division of Medical Sciences, School of Molecular and Clinical Medicine, University of Edinburgh, Edinburgh, UK. [38]Department of Pediatrics, Nemours Children's Hospital, Orlando, Florida, USA. [39]Department of Pathology and Lab Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA. [40]Departments of Pediatrics and Human Genetics, McGill University Health Centre Research Institute, Montréal, Québec, Canada. [41]Division of Gastroenterology, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. [42]Division of Pulmonary Medicine, The Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA. [43]Present address: Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, Illinois, USA. Correspondence should be addressed to H.H. (hakonarson@email.chop.edu).

ONLINE METHODS

Study population. Affected subjects and controls were identified either directly as described in prior studies[61–70] or from de-identified samples and associated electronic medical records (EMRs) in the genomics biorepository at The Children's Hospital of Philadelphia (CHOP). The predominant majority (>80%) of the included cases for IBD, T1D and CVID have been described in previous publications.

Details of each study population are outlined below. EMR searches were conducted with previously described algorithms based on phenotype mapping established using phenome-wide association study (PheWAS) ICD-9 code mapping tables[61–63,70] in consultation with qualified physician specialists for each disease cohort. All DNA samples were assessed for quality control (QC) and genotyped on the Illumina HumanHap550 or HumanHap610 platform at the Center for Applied Genomics (CAG) at CHOP. Note that the patient counts below refer to the total recruited sample size from which we excluded non-qualified samples or genotypes that did not pass QC criteria required for inclusion in the genetic analysis (for example, because of relatedness or poor genotyping rate).

The IBD cohort comprised 2,796 individuals between the ages of 2 and 17, of European ancestry, and with biopsy-proven disease, including 1,931 with CD and 865 with UC and excluding all patients with unclassified IBD. Affected individuals were recruited from multiple centers from four geographically discrete countries and were diagnosed before their 19th birthday according to standard IBD diagnostic criteria, as previously reported[63,65].

The T1D cohort consisted of 1,120 subjects from nuclear family trios (one affected child and two parents), including 267 independent Canadian T1D patients collected in pediatric diabetes clinics in Montreal, Toronto, Ottawa and Winnipeg and 203 T1D patients recruited at CHOP since September 2006. All patients were Caucasian by self-report and between 3 and 17 years of age, with a median age at onset of 7.9 years. All patients had been treated with insulin since diagnosis. Disease diagnosis was based on these clinical criteria, rather than on any laboratory tests.

The JIA cohort was recruited in the United States, Australia and Norway and comprised a total of 1,123 patients with onset of arthritis at less than 16 years of age. JIA diagnosis and JIA subtype were determined according to the International League of Associations for Rheumatology (ILAR) revised criteria[71] and confirmed using the JIA Calculator[72] (http://www.jra-research.org/JIAcalc/), an algorithm-based tool adapted from the ILAR criteria. Prior to standard QC procedures and exclusion of non-European ancestry, the JIA cohort comprised 464 subjects of self-reported European ancestry from Texas Scottish Rite Hospital for Children (Dallas, Texas, USA) and the Children's Mercy Hospitals and Clinics (Kansas City, Missouri, USA); 196 subjects from CHOP; 221 subjects from the Murdoch Children's Research Institute (Royal Children's Hospital, Melbourne, Australia); and 504 subjects from Oslo University Hospital (Oslo, Norway).

The CVID study population consisted of 223 patients from Mount Sinai School of Medicine (MSSM; New York, New York, USA), 76 patients from University of Oxford, (London, England), 47 patients from CHOP, and 27 patients from University of South Florida (USF; Tampa, Florida, USA). The diagnosis in each case was validated against the ESID-PAGID diagnostic criteria, as previously described[73]. Although the diagnosis of CVID is most commonly made in young adults (ages 20–40), all of the CHOP and USF subjects had pediatric-age-of-onset disease, whereas the majority of the subjects from MSSM and Oxford had onset in young adulthood. We note that as the number of individuals with adult-onset CVID is so small (less than 5% of all cases presented) and all ten diseases studied here can present with pediatric age of onset, we elected to refer to the cohort material as pAID.

Q14 The balance of the pediatric subjects' (THY, ~~SPA,~~ PSOR, CEL and SLE) samples were derived from our biorepository at CHOP, which includes more than 50,000 pediatric patients recruited and enrolled by CAG at CHOP Q15 (**Supplementary Table ~~14~~** includes details of genotyped subjects within the CAG pediatric biobank). These individuals were confirmed for diagnosis of THY, SPA, PSOR, CEL and SLE in the age range of 1–17 years at the time of diagnosis and were required to fulfill the clinical criteria for these respective disorders, as confirmed by a specialist. Only patients that upon EMR search were confirmed to have at least two or more in-person visits, at least one of which was with the specified ICD-9 diagnosis code(s), were pursued for clinical confirmation (**Supplementary Table ~~15~~** presents ICD-9 inclusion and exclusion codes). We used ICD-9 codes previously identified and used for PheWASs or EMR-based GWASs and agreed upon by board-certified physicians[62,63]. Q16

Age- and gender-matched control subjects were identified from the CHOP-CAG biobank and selected by exclusion of any patient with any ICD-9 codes for disorders of autoimmunity or immunodeficiency[61] (http://eicd9.com/). Research ethics boards of CHOP and other collaborating centers approved this study, and written informed consent was obtained from all subjects (or their legal guardians). Genomic DNA extraction and sample QC before and after genotyping were performed using standard methods as described previously[64]. All samples were genotyped at CAG on HumanHap550 and 610 BeadChip arrays (Illumina, CA). To minimize confounding due to population stratification, we included only individuals of European ancestry (as determined by both self-reported ancestry and principal-component analysis (PCA)) for the present study. Details of the PCA are provided below.

Genotyping, imputation, association testing and QC. *Disease-specific QC.* We merged the genotyping results from each disease-specific cohort with data from the shared controls before extracting the genotyping results from SNPs common to both Infinium HumanHap550 and 610 BeadChip array platforms and performing genotyping QC. SNPs with a low genotyping rate (<95%) or low MAF (<0.01) or those significantly departing from the expected Hardy-Weinberg equilibrium (HWE; $P < 1 \times 10^{-6}$) were excluded. Samples with low overall genotyping call rates (<95%) or determined to be of outliers of European ancestry by PCA (>6.0 s.d. as identified by EIGENSTRAT[74]) were removed. In addition, one of each pair of related individuals as determined by identity-by-state analysis (PI_HAT > 0.1875) was excluded, with cases preferentially Q17 retained where possible.

Merged-cohort QC. To prepare for whole-genome imputation across the entire study cohort, we combined case samples across the 10 pAIDs with the shared control samples. We repeated the genotyping and sample QC with the same criteria as described above, leaving a final set of ~486,000 common SNPs passing individual-cohort and merged-cohort QC. We again performed identity-by-state analysis and removed related samples (in order to remove related subjects that may have been recruited for different disease studies). We also repeated the PCA and removed population outliers. The final cohort, after the application of all QC metrics mentioned above, included a total of 6,035 patients representing ten pAIDs and 10,718 population-matched controls.

Note that because of the merged QC, compared with the sum of all ten disease-specific GWASs, the final case and control counts in the merged cohort were smaller than the "sum of all cases and controls" (~~**Fig. 1a**~~). In addition, to avoid the potential for confounding due to the presence of duplicated samples, we assigned individuals fitting the diagnostic criteria for two or more pAIDs to whichever disease cohort had the smaller (or smallest) sample size. No subject was included twice. A total of 160 subjects in the study cohort fulfilled criteria for two or more diseases but were counted only once in our reported total of 6,035 unique subjects.

Whole-genome phasing and imputation. We used SHAPEIT[75] for whole-chromosome prephasing and IMPUTE2 (ref. 76) for imputation to the 1KGP-RP (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html, June 2014 haplotype release). For both, we used parameters suggested by the developers of the software and described elsewhere~~to account~~[75–77]. Imputation was done for each 5-Mb regional chunk across the genome, and data were subsequently merged for association testing. Prior to imputation, all SNPs were filtered using the criteria described above.

To verify the imputation accuracy, we validated randomly selected SNPs that reached a nominally significant *P* value after imputation. Because commercially designed genotyping probes were not readily available, we performed Sanger sequencing by designing primers to amplify and sequence the 200-bp region around the imputed SNP markers for two separate 96-well plates. We manually visualized and examined sequences and chromatograms using SeqTrace[78]. Results from this are presented in **Supplementary Table ~~2a~~**, showing >99% mean accuracy.

In addition, a subset of the IBD and CVID subjects were subsequently genotyped on the Immunochip (Illumina) platform. We compared the genotype concordance of all pAID GWAS imputed SNPs that were directly genotyped on

doi:10.1038/nm.3933

the Immunochip after performing sample and marker QC as described above. Results are shown in **Supplementary Table ~~2b.~~**

Disease-specific association testing. We performed whole-genome association testing using post-imputation genotype probabilities with the software SNPTEST (v2.5)[24]. We used logistic regression to estimate odds ratios and betas, 95% confidence intervals and *P* values for trend, using additive coding for genotypes (0, 1 or 2 minor alleles). For autosomal regions, we used a score test, whereas for regions on ChrX we used the ChrX-specific SNPTEST method Newml. QC was performed directly after association testing, excluding any SNPs with an INFO score of <0.80, HWE $P < 1 \times 10^{-6}$, and MAF < 0.01 (overall).

In all analyses, we adjusted for both gender and ancestry by conditioning on gender and the first ten principal components derived from EIGENSTRAT PCA[79]. The λ_{GC} values for all cohorts were within acceptable limits; the highest was observed for the cohort with the largest case sample size, namely, CD ($\lambda_{GC} < 1.07$), consistent with what was previously reported for this data set[65]. In fact, we have previously reported on all the non-CHOP cases included in the present analysis in individual studies using CHOP controls and shown that these individual case-control analyses were well controlled for genomic inflation[61–70]. A QQ plot is provided for each independent cohort in **Supplementary Figure 2a**.

Meta-analysis to identify shared pAID association loci. To identify association loci shared across pAIDs, we meta-analyzed the summary-level test statistics from each of the study cohorts after extracting those markers that passed post-association testing QC for all ten individual disease-specific analyses. To adjust for confounding due to the use of a shared or pooled control population, we applied a previously published method to perform an inverse weighted χ^2 meta-analysis[80].

We LD-clumped the results of the meta-analysis (PLINK) and identified 27 LD-independent associations ($r^2 < 0.05$ within 500 kB up- or downstream of the lead or most strongly associated SNP) reaching a conventional genome-wide significance threshold of $P_{META} < 5 \times 10^{-8}$. We observed that the calculated meta-analysis λ_{GC} was less than 1.09. As recently discussed by de Bakker and colleagues and shown in a number of large-scale GWAS publications, λ_{GC} is related to sample size[81]. As discussed by Yang *et al.*, λ_{GC} depends on the relative contribution of variance due to population structure and true associations versus sampling variance: with no population structure or systematic error, inflation would still depend on heritability, genetic architecture and study sample size[82]. On the basis of de Bakker *et al.*'s recommendations, we also calculated a sample-size-adjusted λ_{1000} by interpolating the λ_{GC} that would have been expected if this study had included only 1,000 cases and 1,000 controls. We performed this only for the meta-analysis results, as the case and control counts for the meta-analysis were both significantly greater than 1,000 (**~~Fig. 1a~~**).

Model search to identify pAIDs associated with the lead signals. The meta-analysis identified SNPs significantly associated with at least one pAID. To determine which pAIDs each SNP was most strongly associated with, we performed a model or 'disease-combination' search. For the lead SNP in each pAID-association locus, we searched for the pAID disease combination that, when the corresponding cases were merged in a mega-analysis, yielded the largest association test statistic.

To identify the disease phenotypes most likely contributing to each identified association signal, we applied the "h.types" method as implemented in the R statistical software package ASSET[83] to perform an exhaustive disease-subtype model search. Note that ASSET provides both a method for genotype-level association testing (h.types used in this study) and a summary-level modified fixed-effect meta-analysis approach ("h.traits") that allows for heterogeneity of SNP effects across different phenotypes. Both methods exhaustively enumerate each combination of phenotypes that are jointly considered, and therefore test a total of

$$\text{pAID disease model combinations} = \sum_{i=1}^{r} \frac{n!}{(n-i)!(i)}$$

where *r* is the total number of disease subtypes assigned to cases (for example, ranging from one to ten pAIDs) and *n* is the total number of disease subtypes (i.e., ten pAIDs). Note that this reduces to $2^n - 1$ (or 1,023 unique combinations here), as in this case we considered all possibilities of *r* across *n* of ten diseases. The ASSET algorithm iteratively tests each pAID case combination using logistic regression to determine whether there is an association between genotype counts and case status. For each SNP tested, the 'optimal' subtype model is the combination of pAIDs that, when tested against the shared controls in the logistic regression analysis.

Identification of lead associated variants showing opposite direction of effect. For each of the top 46 associating loci ($P_{META} < 1 \times 10^{-6}$), we identified those loci for which the lead SNP had an effect direction (on the basis of logistic regression betas) opposite that reported for the disease combination identified by the subtype model search and whose corresponding association *P* value reached at least nominal significance ($P < 0.05$). We identified nine instances.

Candidate gene prioritization. To annotate the lead SNPs to candidate genes, we prioritized the mapping to candidate genes systematically in the following manner:

1. If the SNP or locus was previously reported in autoimmune diseases at genome-wide significance, we provided the candidate gene symbol, where available, as identified in the GWAS Catalog[84] or ImmunoBase[83].
2. If an SNP was annotated as coding or fell within the coding DNA sequence (i.e., intronic or in the UTRs), we reported that gene as identified by the variant effect predictor (VEP)[85].
3. If the SNP was upstream, downstream, or intergenic, we prioritized the gene by using the best candidate gene identified with the network tool DAPPLE[86].
4. If none of the above was feasible, we manually curated the most 'likely' gene on the basis of the observed LD block and evidence of prior association signals with autoimmune diseases or other immune-related phenotypes as presented in the dbSNP or GWAS catalog.

Functional or biological annotations and enrichment analysis using publicly accessible resources. We annotated the lead pAID-associated SNPs using publicly available functional and biological databases and resources. We considered the top imputed lead SNP for each locus and, in addition, any of its near-perfect proxies (defined as $r^2 > 0.8$ within 500 kB up- or downstream) on the basis of the 1KGP-RP.

We included annotation, expression, interaction and network data from the following resources:

1. Genomic mapping and annotation: SNAP[87], SNP-Nexus[88], Ensemble[89] and UCSC[90].
2. Regulatory annotations: EnCODE (TF-binding sites and DNase-hypersensitivity sites)[91], GTex[92] (eQTLs), and a published lymphoblastoid cell line eQTL data set[93].
3. Functional annotations: SIFT[94], Polyphen[95], miRNA target site polymorphisms[96,97].
4. Conservational or evolutionary predictions: GERP[98], PHAST++[99], CpG islands[100].
5. Literature search: GAD[101], NHGRI GWAS catalog[102], dbGAP[103], or published Immunochip studies[104] (www.immunobase.org) for literature support.
6. Gene expression and enrichment analysis: ImmGen[102] (murine) and whole-transcriptome analysis across 126 tissues[104] (human).
7. Protein-protein interaction (PPI) database: DAPPLE[86], STRING[105].
8. Pathway-based and gene set enrichment analysis: Gene Ontogeny[106], Webgestalt[107], Wikipathways[108], IPA[109], DAVID[110], GSEA[111], and Pathways Commons[112].
9. Gene network analysis and visualization: DAPPLE[86] and VEP[85] to prioritize candidate causal genes and Grail[113] for text-mining of PubMed database for coassociations.

Functional and biological annotations (categories 1–5) for the 27 lead SNPs are illustrated in **Figure 3**; annotations are also provided for the 46 GWM loci in **Supplementary Figure ~~4~~**. The following annotation types were used:

1. Regulatory: EnCODE consensus TF-binding sites (T), DNase I hypersensitivity sites (S), or published eQTL signals (E)

2. Functional: known mutations in PolyPhen or SIFT (A), experimentally validated (miRBASE 18.0) and predicted (mirSNP) miRNA target sites (R), or SNPs that tag regions containing common copy-number variation regions reported by the database of genomic variants (DGV) (V)
3. Conserved: conserved nucleotide sequences based on GERP++/phastCon (C) or known CpG islands that correlate with epigenetic methylation patterns (M)
4. Literature-supported: published association with immune or inflammatory diseases or immune-related endophenotypes from candidate studies or GWASs catalogued in the Genetic Association Database, NHGRI GWAS catalog, dbGAP, or Immunochip studies (L)

In addition to determining whether the 27 GWS pAID-associated SNPs were enriched for a given annotation type, we performed Monte Carlo simulations to resample 10,000 times the SNPs (MAF > 0.01 in Europeans) from all SNPs in 1KGP-RP. As for the 27 lead SNPs, for each set of 100 randomly sampled SNPs, we expanded the list by first identifying all nearby SNPs in strong LD (i.e., LD proxies with $r^2 > 0.8$ within 500 kB up- or downstream) within the 1KGP-RP data set filtered for only SNPs with MAF > 0.01 in the European population. We then annotated each original and any proxy SNPs as above for each major annotation category. We collapsed the information for all proxies identified for a given lead such that for any given category, if the lead SNP or any of its proxies were annotated, the lead SNP was marked as annotated. We then calculated the frequency of annotation for the 100 SNPs in each set. After sampling and annotating 100-SNP sets 10,000 times, we use the permutation-derived distribution of annotation percentages for each annotation type to calculate an enrichment *P* value such that

$$P_{\text{enrich}} = 1 - \frac{\text{Count}(f_{\text{pAID}} > F_{\text{conclusive}})}{10{,}000}$$

where *f* is the percentage of SNPs in the pAID set that are annotated and *F* is the distribution of the percentage of SNPs annotated across 10,000 sets of 100 SNPs resampled from the 1KGP-RP using only markers with MAF > 0.01 in Europeans.

Hierarchical clustering based on effect size and direction of association. We performed agglomerative hierarchical clustering across the top 27 independent loci using the directional *Z*-score obtained from logistic regression analysis in each of the ten disease-specific GWASs, defined as

$$Z = \frac{\text{beta (effect size)}}{\text{s.e.}}$$

The standardized and normalized *Z*-scores were used as inputs to the agglomerative hierarchical clustering. We used Ward's minimal-variance method to identify relatively consistent gene and locus cluster sizes.

Gene-based association testing. Given our interest in genetic overlap across pAIDs, we sought to identify genes associated with pAIDs in a disease-agnostic manner that was insensitive to locus and phenotypic heterogeneity. We used VEGAS[114], a set-based method, to perform GBAT.

As input, we used the nominal P_{META} values from the pooled, inverse χ^2 meta-analysis for the ten pAIDs across the genome as the input summary statistics for VEGAS, without considering which specific diseases were identified in the model search analysis. We assigned SNPs to gene regions and performed 10^7 simulations to estimate the gene-based *P* value as described in VEGAS's documentation. We used two thresholds: $P_{\text{sim}} < 2.8 \times 10^{-6}$ to identify significant candidate genes, on the basis of a Bonferroni adjustment for approximately 17,500 genes tested, and a false discovery rate (FDR) of <2%, which corresponds to a *q* value of <0.0205, which was used only for pathway and gene set enrichment analysis.

Tissue-specific gene set enrichment analysis. With few exceptions, most genes that are known to have a causative role in autoimmune disease have been shown to regulate molecular or subcellular processes in immune or immune-related tissues. If candidate pAID-associated genes are relevant to autoimmune-disease biology, then expression of these genes would be expected to be, on average, higher across immune or immune-related tissues (as compared with expression in non-immune-related tissues). Thus, we compared the expression of candidate pAID-associated genes identified by GBAT with that of non-candidate genes in a variety of tissues.

We curated the expression of the transcriptome in a broad spectrum of human tissues using a publicly available data set consisting of summary-level, normalized gene expression levels for more than 12,000 unique genes across 126 tissues and/or cell types, including a large number of immune tissues and cells[104]. We downloaded the processed data set "mean expression data matrix."

Across the 126 unique tissues, we tested whether the median or cumulative distribution of expression of pAID-associated gene transcripts as identified by GBAT was higher than that of the remaining transcripts in the data set using a one-sided Wilcoxon rank test or a one-sided Kolmogorov-Smirnov (KS) test, respectively. We calculated a tissue-specific gene expression ES value, which is the $-\log_{10}$ (*P* value) obtained from comparing the relative enrichment in transcript expression of pAID-associated genes versus the transcripts of the remaining genes in the data set. The tests were done on a per-tissue basis to derive a set of KS and a set of Wilcoxon ES values. We performed this per tissue analysis (1) for the total set of pAID-associated genes from GBAT and (2) when genes across the extended MHC (chr6: 25–34 Mb) were excluded.

We performed the secondary immune–versus–non-immune comparative analysis by plotting the ES values obtained from either Wilcoxon or KS tests in descending rank order of the respective test statistics, as shown in **Figure ~~3a~~** and **Supplementary Figure ~~5~~** for all 126 tissue types. In those figures each point represents a single tissue and is colored according to its classification as either immune (red) or non-immune (blue), as described previously[86]. To formally test whether the overall ES values were higher among immune tissues than among non-immune tissues, we performed both the Wilcoxon rank sum test and the KS test on the vector of per-tissue ES values, comparing those derived from immune and non-immune tissues. We found that the enrichment observed across immune tissues was specific and not general to any GWAS-identified signals. We repeated this analysis in two sets of candidate genes, one for CD and another for schizophrenia, by identifying all associated genes for the two phenotypes from the NHGRI GWAS Catalog.

Immune cell gene set enrichment analysis. Cells of the immune system are extremely diverse in function and gene expression. To more precisely assess the expression of pAID-associated genes, we examined the mRNA expression of pAID candidate genes across specific immune cell subtypes, as well as during different developmental time points.

ImmGen provides a publicly available, high-quality murine gene expression data set. The ImmGen data set consists of 226 murine immune cell types across different lineages at multiple developmental stages, sorted by FACS and assayed at least in triplicate. Standard QC and quantile-normalization methods were applied to the data set as described by ImmGen[102]. The total set of transcripts mapped to 14,624 homologs in the human transcriptome on the basis of genes annotated in the hg18/build36 of the human reference genome, which were used to query the gene expression data.

Some of the cell types were derived from genetically altered animals, and the results from analysis of those cell types would have been difficult to interpret, so we removed those cell lines from the analysis. The complete list of cell types used in the analysis and the category to which we assigned each cell type for the categorical analysis are presented in **Supplementary Table ~~8b~~**. A total of 176 unique cell lines remained for subsequent analyses using this data set.

As with the human data set, we calculated the ES values by comparing the expression of the pAID-associated candidate gene transcripts to that of the remaining transcripts assayed in the data set for each immune cell type examined. We plotted the distribution of relative gene expression ES values as a density plot across the range of ES values from all of the examined cell types available. We compared the results obtained using the full set of candidate pAID genes identified by GBAT or obtained when we excluded the genes within the extended MHC. To ensure that this was not simply a result of selection bias (as GWASs may be biased toward regions or genes across the genome that are better sampled or more densely genotyped), we compared the results to those obtained with the curated gene lists from the GWAS catalog (as above) for CD, schizophrenia, body mass index and LDL cholesterol.

To determine whether pAID-associated candidate genes are expressed at higher levels (relative to the rest of the genes in the transcriptome) in some

doi:10.1038/nm.3933

immune cell types than in others, we defined immune cell types according to surface marker expression and tissue isolation details provided by ImmGen. Some categories were further divided into subcategories (for example, B and T cells) on the basis of developmental stage or lineage into a total of 16 non-overlapping cell-type categories. To compare the results across the cell-type categories, we plotted the distribution of ES value ranks for each cell type, binning the results according to the category each cell type belonged to (again, we performed the analysis either with or without the extended MHC region).

Expression profiling of pleiotropic autoimmune disease–associated genes across specific immune cell types. We profiled the expression of genes that had been identified in at least three autoimmune diseases in our subtype model search, previously published Immunochip fine-mapping studies, or a combination thereof (for example, identified as associated with JIA and UC in our analysis but previously identified as a candidate gene from an Immunochip analysis of ~~AA~~). We identified 217 candidate pleiotropic genes, of which 191 could be mapped to unique gene transcripts within the ImmGen data sets.

We performed agglomerative hierarchical clustering with the matrix of gene expression levels from the 191 candidate gene transcripts using Ward's minimal-variance method across all 176 immune cell types. The genes and cell types shown in dendrograms are based on the results of unsupervised hierarchical clustering analysis and represent four major groups of cells and six major groups of genes.

We examined whether genes that were clustered on the basis of similar immune cell–expression profiles were likely to be associated with the same disease(s). Specifically, given a set of genes associated with one or more autoimmune diseases grouped in cluster i (C_i), we asked whether there is an increased likelihood (i.e., more so than expected by chance as compared with genes not found within this cluster) that these genes are also associated with disease j (D_j), such that

	C_i (yes)	C_i (no)
D_j (yes)	a	b
D_j (no)	c	d

where the expected probability of the values observed under the null is given by the hypergeometric distribution. As some of the cell counts were small and we were interested only in identifying instances where $a >> b$, c or d, we used a one-sided Fisher's exact test. We first tested each of the 18 autoimmune diseases across all identified clusters, declaring nominal and Bonferonni-adjusted significance at $P < 0.05$ and $P < 5.6 \times 10^{-4}$, respectively. For any clusters where at least two diseases reached nominal or marginal significance, we also tested whether there was an overrepresentation of genes associated with both diseases at $P < 0.05$.

PPI and network analysis. DAPPLE[86]: PPIs among the set of either 27 GWS or 46 GWM candidate regions were identified; the input seeds were defined as the 100-kB sequences up- and downstream of the most significantly associated SNP (based on hg19) in each candidate region. Other input parameters included 50-kB regulatory region length, a common interactor binding degree cutoff of 2, and the following specified known genes: *IL23R*, *PTPN22*, *INS*, *NOD2*, *DAG1*, *SMAD3*, *ATG16L1*, *ZNF365*, *PTGER4*, *NKX2-3*, *ANKRD55* and *IL12B*. We performed 10,000 permutations to accurately calculate enrichment network statistics. Seed scores P_{dapple} were used to color the protein nodes in the network plot.

STRING[105]: We used the *Homo sapiens* PPI database to query one of three lists: (1) the GWS loci, (2) GWS and GWM loci or (3) the list of genes identified by GBAT shown to be enriched for key proteins in the JAK-STAT pathway. We assessed and reported the evidence of PPI enrichment on the basis of these queries as compared to the results expected for the rest of the genes in the human genome. We generated network plots for the directly connected protein candidates (~~the~~ **Supplementary** ~~Figures~~ represent the "evidence" plot option).

Pathway and gene set enrichment analysis. Webgestalt[107]: For pathway and gene set analysis, we used the web-based tool Webgestalt to examine evidence of shared TF binding, miRNA target–binding sites, and enrichment in specific Gene Ontology and Pathway Commons categories. The inputs for this analysis included all lead genes (FDR < 2%) from the GBAT (similar to that for the other pathway annotation databases below for consistency).

DAVID[110]: We used the bioinformatics web tool DAVID (v6.7, available at http://david.abcc.ncifcrf.gov) for functional-annotation analysis of the significant genes. Significant genes with FDR < 2% in VEGAS, the gene-based association analysis, were used as input for DAVID. DAVID performed overrepresentation analysis of functional-annotation terms on the basis of hypergeometric testing and adjusted for multiple testing. To compare the results of this analysis with results obtained via other methods, we used BioCarta, KEGG pathways and GO_BP_FAT as gene set definition files.

IPA[109]: We used IPA software (http://www.ingenuity.com/) for canonical pathway and network analysis. We inputted all the significant genes in the VEGAS output (FDR < 2%) for IPA analysis. In the IPA core analysis, we selected the Ingenuity Knowledge Base (Genes Only) as the reference set, including both direct and indirect relationships. We used the filter setting of relationships in human and experimentally observed only. Information regarding canonical pathways was obtained from IPA output.

GSEA[115,116]: We conducted gene set enrichment analysis with the software GSEA (http://www.broadinstitute.org/gsea) using as input the pre-ranked gene list generated on the basis of the –log(P value) from VEGAS using all genes. We selected the following settings for our analysis: number of permutations, 5,000; enrichment statistic, weighted; maximum size of gene set, 500; minimum size of gene set, 15; and with normalization.

Interdisease genetic sharing analysis. To examine the degree of overlap in genetic risk susceptibility between any two autoimmune diseases, we developed and/or implemented the following statistical measures to quantify interdisease genetic sharing:

1. LPS test, optimized to evaluate whether two pAIDs share more loci in common than would be expected to occur by chance; the score 'penalizes' disease pairs if many of the loci are disease specific. The test is helpful if only data on whether diseases share specific candidate genes or association loci in common are known.
2. GPS test, optimized to assess the correlation between the set of association test statistics observed genome-wide across any two pAIDs. This test is valuable because it is independent of the gene sets chosen and thus does not require the use of any arbitrary method to define a significance 'threshold' of input data.

LPS analysis. To quantify the similarity between any two diseases D_1 and D_2 on the basis of the degree to which D_1 and D_2 share independent genetic risk associations (i.e., loci, SNPs or candidate genes), we considered the following model.

We began with a list of candidate genes, association loci or LD-independent SNPs n_r identified as having reached a predefined GWAS significance threshold (e.g., GWS or GWM) across one or more SNPs from n_r for a set of diseases with expected or hypothesized sharing (i.e., all autoimmune diseases in this study and those reported on by the Immunochip studies catalogued by ImmunoBase[83]).

For any two diseases D_1 and D_2, a given candidate gene or SNP x_i could be uniquely classified in one of four ways: associated with D_1 and D_2 (n_{11}), associated only with D_1 (n_{12}) or D_2 (n_{21}), or associated with neither D_1 nor D_2 (n_{22}). For any given list of TOP associations (i.e., n_r), the distribution across the four possible categories can be tabulated as follows:

Locus x_i	D_2 (yes)	D_2 (no)
D_1 (yes)	n_{11}	n_{12}
D_1 (no)	n_{21}	n_{22}

where $n_{11} + n_{12} + n_{21} + n_{22} = n_r$ and D_1 (yes) or (no) means the SNP x_i is or is not associated with that marker, respectively.

The probability P_x that an SNP x_i from the list n_r is associated with either D_1 or D_2 can be expressed as

$$P_1 = \frac{n_{11} + n_{12}}{n_r} \quad \text{(for } D_1\text{)}$$

$$P_2 = \frac{n_{12} + n_{21}}{n_r} \quad \text{(for } D_2\text{)}$$

for any two pAIDs D_1 and D_2.

Thus, the frequency at which x_i should truly be associated with two distinct disease subtypes is given by $n_r(P_1P_2)$, and the observed number of overlapping associations is represented by n_{11}. Therefore, under the null hypothesis H_0, for a given pair of diseases D_1 and D_2, the variance of the difference between the numbers of expected and observed associations of all those tested (n_T) shared by both D_1 and D_2 should follow a normal distribution.

$$Z = \frac{n_{11} - n_r(P_1P_2)}{\sqrt{n_r(P_1P_2)(1-P_1P_2)}} \sim N(0,1)$$

We used the one-sided Z-test to examine whether the degree of overlap was significantly greater than expected, assuming a normal distribution under the null hypothesis that D_1 and D_2 do not share more associations than they would by chance. We used a Bonferroni adjustment to correct for 45 pairwise disease-combination tests.

GPS analysis. The GPS test determines whether two pAIDs are genetically related. For the ith SNP, let $X_i = 1$ if the SNP is truly associated with one disease, and let $X_i = 0$ otherwise. Similarly, define Y_i as the indicator of whether the SNP is associated with the other disease in the pair. We can therefore consider the diseases to be genetically related if there are more SNPs with $(X_i,Y_i) = (1,1)$ than would be expected to occur by chance. This amounts to testing the independence of X_i and Y_i.

However, we do not directly observe X_i and Y_i and instead observe P values U_i and V_i, which come from the two GWAS studies for the two diseases. When $X_i = 1$, the P value U_i will tend to be small, and otherwise U_i will be uniformly distributed; the same is true of Y_i and V_i. If U_i and V_i are independent, then X_i and Y_i must be as well. We can therefore test for genetic relatedness by testing whether the P values are dependent.

Most existing methods may not take advantage of the availability of the full genome data set for testing genetic sharing using U_i and V_i. To address this limitation, we developed a novel, threshold-free method to detect genetic relatedness. Our test statistic is defined by

$$D = \sup_{u,v} \sqrt{\frac{n}{\ln n}} \frac{\left|F_{uv}(u,v) - F_u(u)F_v(v)\right|}{\sqrt{F_u(u)F_v(v) - F_u(u)^2F_v(v)^2}}$$

where n is the total number of SNPs, $F_{uv}(u,v)$ is the empirical bivariate distribution function of (U_i,V_i), and $F_u(u)$ and $F_v(v)$ are the empirical univariate distribution functions of U_i and V_i, respectively. Intuitively, the numerator of D is motivated by the fact that if U_i and V_i are truly independent, their bivariate distribution is equal to the product of their univariate distributions. The denominator of D makes the test capable of detecting even very weak correlations. It can be shown that D is asymptotically optimal for testing for genetic relatedness. Under the null hypothesis of no genetic sharing, it can be shown that D is approximately distributed like the inverse square root of a standard exponential random variable. This gives us an analytic expression for calculating P values. Note that no significance threshold is required.

The asymptotic null distribution of D is derived under the assumption that the genetic markers examined across the genome are statistically independent. We therefore pruned the SNPs for each pair of diseases before applying our test. We conducted inverse χ^2 meta-analyses separately for each pair of diseases and pruned the resulting P values using a threshold of $r^2 < 0.5$ within a 500-kB up- and downstream region. This left about 800,000 SNPs for each disease pair analyzed. The use of more stringent r^2 thresholds (for example, $r^2 < 0.3$ or 0.2) gave comparable results.

Because the GPS test has the underlying assumption that the genetic markers examined across the genome are statistically independent, we applied the test to independent P values for every pairwise pAID combination examined using a threshold of $r^2 < 0.5$ within a 500-kB up- and downstream region. This left about 800,000 SNPs for each disease pair analyzed. Note that we observed similar results when we used either the full genome-wide data set or a reduced marker set with more stringent r^2 thresholds (for example, $r^2 < 0.3$ or 0.2).

Undirected weighted cyclic network visualization of results from the locus-specific sharing test. In graphic representations, pairwise relationships between autoimmune diseases (nodes) are represented by edges, whose weights are determined by the magnitude of the LPS test statistic (R statistical software package q-graph). Specifically, the width and density of the edges are the standardized transformations of the test statistic, and the colors denote whether the direction of the test statistic is positive (blue, meaning more sharing than expected) or negative (red, meaning less sharing than expected). Although graphs are constructed from all 45 pairwise interactions, for simplicity and improved visualization, we showed only those edges that represented a pairwise interaction that reached a Bonferroni-adjusted (**Supplementary Fig. ~~6b~~**) or nominal (**Supplementary Fig. ~~7~~**) significance threshold ($P < 0.05$). The nodes are positioned on the basis of a force-directed layout based on the Fruchterman-Reingold algorithm.

In silico replication of novel pAID-association loci using previously published autoimmune disease cohort data sets. Replication set I: The following data sets were used in the first replication set ~~(Table 1).~~ CASP[117], CIDR Celiac Disease[118], NIDDK Crohn's Disease[119], Wellcome Trust Case Control Consortium (WT) Crohn's Disease and Type 1 Diabetes[120], WT Ulcerative Colitis[121] and WT Ankylosing Spondylitis[122]. These data sets were obtained via dbGaP or the Wellcome Trust Case Control Consortium. In order to maximize the power, we ~~tried to replicate~~ each of the 12 significant SNPs in all of the seven available data sets.

Each data set was subjected to strict QC filtering as follows: we removed individuals that were inferred to be related on the basis of genetic data, individuals with >10% missing data, individuals with a reported sex that did not match the observed heterozygosity rates on chromosome X, and individuals not of European ancestry. We further removed variants with >10% missingness, variants not in HWE, variants with missingness significantly correlated to phenotype, and variants with MAF < 0.005. Variants to be replicated that were not observed in the original data set were imputed using IMPUTE2 (ref. 123) and the 1KGP-RP haplotype data[124]. Markers across the X chromosome, which were previously considered by most of these studies, were reanalyzed using the XWAS toolset[125,126].

Replication-association analysis was carried out by logistic regression implemented in PLINK[127]. The first ten principal components calculated using EIGENSOFT[128] were added as covariates for all data sets except CASP, where no population stratification was observed.

Replication set II: The second replication set consisted of the following data sets: Rheumatoid Arthritis meta-analysis[129], IBDG Ulcerative Colitis meta-analysis[130], IBDG Crohn's Disease meta-analysis[131], Systemic Lupus Erythematosus GWAS[132], and SLEGEN[133]. Individuals from these data sets were of European ancestry. Summary statistics from the original studies were publicly available and were used for the replication analysis. Details regarding QC procedures and association analysis can be obtained from the original studies[129–133].

LD-based replication for replication sets I and II: We further assessed replication in SNPs that were in LD with the significant SNPs in the discovery set. For each associated SNP, a list of SNPs in LD ($r^2 > 0.5$) within 500 kb of the original SNP was obtained from SNAP[87] using the 1KGP-RP. Q23

61. Denny, J.C. *et al.* PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* **26**, 1205–1210 (2010).
62. Ritchie, M.D. *et al.* Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am. J. Hum. Genet.* **86**, 560–572 (2010).
63. Liao, K.P. *et al.* Associations of autoantibodies, autoimmune risk alleles, and clinical diagnoses from the electronic medical records in rheumatoid arthritis cases and non-rheumatoid arthritis controls. *Arthritis Rheum.* **65**, 571–581 (2013).
64. Hakonarson, H. *et al.* A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature* **448**, 591–594 (2007).
65. Imielinski, M. *et al.* Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nat. Genet.* **41**, 1335–1340 (2009).
66. Kugathasan, S. *et al.* Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat. Genet.* **40**, 1211–1215 (2008).
67. Orange, J.S. *et al.* Genome-wide association identifies diverse causes of common variable immunodeficiency. *J. Allergy Clin. Immunol.* **127**, 1360–1367.e6 (2011).
68. Behrens, E.M. *et al.* Association of the TRAF1-C5 locus on chromosome 9 with juvenile idiopathic arthritis. *Arthritis Rheum.* **58**, 2206–2207 (2008).
69. Grant, S.F. *et al.* Association of the *BANK 1* R61H variant with systemic lupus erythematosus in Americans of European and African ancestry. *Appl. Clin. Genet.* **2**, 1–5 (2009).

doi:10.1038/nm.3933

70. Liao, K.P. *et al.* Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res. (Hoboken)* **62**, 1120–1127 (2010).
71. Petty, R.E. *et al.* International League of Associations for Rheumatology classification of juvenile idiopathic arthritis: second revision, Edmonton, 2001. *J. Rheumatol.* **31**, 390–392 (2004).
72. Behrens, E.M. *et al.* Evaluation of the presentation of systemic onset juvenile rheumatoid arthritis: data from the Pennsylvania Systemic Onset Juvenile Arthritis Registry (PASOJAR). *J. Rheumatol.* **35**, 343–348 (2008).
73. Conley, M.E., Notarangelo, L.D. & Etzioni, A. Diagnostic criteria for primary immunodeficiencies. Representing PAGID (Pan-American Group for Immunodeficiency) and ESID (European Society for Immunodeficiencies). *Clin. Immunol.* **93**, 190–197 (1999).
74. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
75. Delaneau, O., Coulonges, C. & Zagury, J.-F. Shape-IT: new rapid and accurate algorithm for haplotype inference. *BMC Bioinformatics* **9**, 540 (2008).
76. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
77. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457–470 (2011).
78. Stucky, B.J. SeqTrace: a graphical tool for rapidly processing DNA sequencing chromatograms. *J. Biomol. Tech.* **23**, 90–93 (2012).
79. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904 (2006).
80. Zaykin, D.V. & Kozbur, D.O. *P*-value based analysis for shared controls design in genome-wide association studies. *Genet. Epidemiol.* **34**, 725–738 (2010).
81. De Bakker, P.I. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **17**, R122–R128 (2008).
82. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
83. Institute for Systems Biology and Juvenile Diabetes Research Foundation–Wellcome Trust Diabetes and Inflammation Laboratory. *ImmunoBase* http://www.immunobase.org (2013).
84. NHGRI. Published GWAS through 08/01/2014. *NHGRI GWA Catalog* http://www.genome.gov/multimedia/illustrations/GWAS_2011_3.pdf (2014).
85. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
86. Rossin, E.J. *et al.* Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet.* **7**, e1001273 (2011).
87. Johnson, A.D. *et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24**, 2938 (2008).
88. Chelala, C., Khan, A. & Lemoine, N.R. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics* **25**, 655–661 (2009).
89. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43**, D662–D669 (2015).
90. Kent, W.J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996 (2002).
91. Boyle, A.P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
92. National Institutes of Health Genotype-Tissue Expression (GTEx) http://commonfund.nih.gov/GTEx/index (2015).
93. Liang, L. *et al.* A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Res.* **23**, 716–726 (2013).
94. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4**, 1073–1081 (2009).
95. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr. Protoc. Hum. Genet.* Chapter 7, Unit 7.20 (2013).
96. Liu, C. *et al.* MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. *BMC Genomics* **13**, 661 (2012).
97. Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. & Enright, A.J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34**, D140–D144 (2006).
98. Davydov, E.V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP. *PLoS Comput. Biol.* **6**, e1001025 (2010).
99. Nguyen, D.-Q. *et al.* Reduced purifying selection prevails over positive selection in human copy number variant evolution. *Genome Res.* **18**, 1711–1723 (2008).
100. Bird, A.P. CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209–213 (1986).
101. Becker, K.G., Barnes, K.C., Bright, T.J. & Wang, S.A. The genetic association database. *Nat. Genet.* **36**, 431–432 (2004).
102. Heng, T.S.P. & Painter, M.W. The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.* **9**, 1091–1094 (2008).
103. Mailman, M.D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**, 1181–1186 (2007).
104. Benita, Y. *et al.* Gene enrichment profiles reveal T-cell development, differentiation, and lineage-specific transcription factors including ZBTB25 as a novel NF-AT repressor. *Blood* **115**, 5376–5384 (2010).
105. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808–D815 (2013).
106. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
107. Wang, J., Duncan, D., Shi, Z. & Zhang, B. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Res.* **41**, W77–W83 (2013).
108. Kelder, T. *et al.* WikiPathways: building research communities on biological pathways. *Nucleic Acids Res.* **40**, D1301–D1307 (2012).
109. Ingenuity Systems Ingenuity Pathway Analysis http://www.ingenuity.com/products/ipa (2015).
110. Huang, D.W. *et al.* The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* **8**, R183 (2007).
111. Wang, K., Li, M. & Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* **81**, 1278–1283 (2007).
112. Cerami, E.G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* **39**, D685–D690 (2011).
113. Raychaudhuri, S. *et al.* Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet.* **5**, e1000534 (2009).
114. Liu, J.Z. *et al.* A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.* **87**, 139–145 (2010).
115. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–15550 (2005).
116. Mootha, V.K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
117. Nair, R.P. *et al.* Genome-wide scan reveals association of psoriasis with IL-23 and NF-κB pathways. *Nat. Genet.* **41**, 199–204 (2009).
118. Ahn, R. *et al.* Association analysis of the extended MHC region in celiac disease implicates multiple independent susceptibility loci. *PLoS ONE* **7**, e36926 (2012).
119. Duerr, R.H. *et al.* A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science* **314**, 1461–1463 (2006).
120. Wellcome Trust Case Control Consortium. *et al.* Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
121. Barrett, J.C. *et al.* Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. *Nat. Genet.* **41**, 1330–1334 (2009).
122. Evans, D.M. *et al.* Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat. Genet.* **43**, 761–767 (2011).
123. Marchini, J. *et al.* A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).
124. Abecasis, G.R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
125. Gao, F. *et al.* XWAS: a software toolset for genetic data analysis and association studies of the X chromosome. *bioRxiv* doi:10.1101/009795.
126. Chang, D. *et al.* Accounting for eXentricities: analysis of the X chromosome in GWAS reveals X-linked genes implicated in autoimmune diseases. *PLoS One* **9**, e113684 (2014).
127. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
128. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
129. Stahl, E.A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42**, 508–514 (2010).
130. Anderson, C.A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43**, 246–252 (2011).
131. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42**, 1118–1125 (2010).
132. Hom, G. *et al.* Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. *N. Engl. J. Med.* **358**, 900 (2008).
133. Harley, J.B. *et al.* Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in *ITGAM*, *PXK*, *KIAA1542* and other loci. *Nat. Genet.* **40**, 204–210 (2008).

QUERY FORM

Nature Medicine	
Manuscript ID	[Art. Id: 3933]
Author	[illegible]
Editor	
Publisher	

AUTHOR:

The following queries have arisen during the editing of your manuscript. Please answer queries by making the requisite corrections directly on the galley proof. It is also imperative that you include a typewritten list of all corrections and comments, as handwritten corrections sometimes cannot be read or are easily missed. Please verify receipt of proofs via e-mail

Query No.	*Nature of Query*
Q1	Please carefully check the spelling and numbering of all author names and affiliations.
Q2	Gene symbols changed throughout when necessary to reflect HGNC-approved nomenclature.
Q3	All tables and table parts must be cited in order. Supplementary Table 1 cited in its entirety here to preclude need to cite Supplementary Table 1b–f before Table Supplementary 2a; OK?
Q4	All figures and figure panels must be cited in order. Supplementary Fig. 1 cited in its entirety here to avoid renumbering of subsequent figure/panel citations; OK?
Q5	Correct that Supplementary Table 1b was meant here, and not Table 1b as originally cited (Table 1 does not seem to have multiple parts)?
Q6	Please cite Supplementary Table 2b before Supplementary Table 2c.
Q7	Please cite Supplementary Fig. 2 before Supplementary Fig. 3.
Q8	Please cite Supplementary Figs. 4–6 before Supplementary Fig. 7, or renumber figures.
Q9	Please cite Supplementary Table 3b before Supplementary Table 3c.
Q10	Please define SJO and SS.
Q11	(1) Should a reference be cited here? (2) Please be more specific about this consortium—the International IBD Genetics Consortium, the Pediatric IBD Consortium, etc.?
Q12	Correct that in the preceding sentence, numbers in brackets after gene symbols in the original were references? If not, please edit for clarity.
Q13	OK as edited?

QUERY FORM

Nature Medicine	
Manuscript ID	[Art. Id: 3933]
Author	Li and
Editor	e
Publisher	

AUTHOR:

The following queries have arisen during the editing of your manuscript. Please answer queries by making the requisite corrections directly on the galley proof. It is also imperative that you include a typewritten list of all corrections and comments, as handwritten corrections sometimes cannot be read or are easily missed. Please verify receipt of proofs via e-mail

Query No.	*Nature of Query*
Q14	Please define SPA.
Q15	There is no Supplementary Table 14; please correct the citation.
Q16	There is no Supplementary Table 15; please correct the citation.
Q17	"Identity-by-state" written out throughout to avoid confusion with irritable bowel syndrome, also abbreviated as IBS elsewhere in the paper.
Q18	All equations were present as embedded images in the original article file and were extremely difficult to read, and errors may have been introduced during re-typing. Please check carefully for accuracy.
Q19	Please check preceding sentence for sense.
Q20	Please cite Supplementary Tables 7 and 8a before Supplementary Table 8b. Also, please note that the Supplementary Material does not indicate multiple parts in Supplementary Table 8.
Q21	Please cite Supplementary Table 9 after Supplementary Table 8.
Q22	Please define AA.
Q23	Some references from the primary reference list are repeated in the online-only reference list. Each reference should have only one number; if references from the primary reference list are cited in the Online Methods, they should be cited with their original numbers, not as new references. Please check both lists carefully and delete duplicate references from the Online Methods, renumbering the list and the citations in the text as needed.
Q24	Please check that all funders have been appropriately acknowledged and that all grant numbers are correct.

QUERY FORM

Nature Medicine	
Manuscript ID	[Art. Id: 3933]
Author	
Editor	
Publisher	

AUTHOR:

The following queries have arisen during the editing of your manuscript. Please answer queries by making the requisite corrections directly on the galley proof. It is also imperative that you include a typewritten list of all corrections and comments, as handwritten corrections sometimes cannot be read or are easily missed. Please verify receipt of proofs via e-mail

Query No.	*Nature of Query*
Q25	Please provide a brief statement of each author's contribution to the work (e.g., A.A. designed the study; B.B. analyzed data; etc.).
Q26	Link in ref. 12 leads to published GWAS through 9/2011, but reference title refers to GWAS through 8/2014. Please check and correct as needed.
Q27	Please cite Fig. 2b before Fig. 3 or Fig. 2 as a whole.
Q28	the original version of this table had PS instead of PSOR. We have changed all of these instances. Check that this is correct. We also changed the height of the columns to ensure they were all uniform. Is this OK?]
Q29	It leads to error page. Please check
Q30	It leads to error page. Please check